Michael Beißwenger

14 Internetbasierte Kommunikation und Korpuslinguistik: Repräsentation basaler Interaktionsformate in TEI

Abstract: Der Beitrag beschreibt ein Basisschema für die Repräsentation von Korpora internetbasierter Kommunikation auf der Grundlage der *Guidelines for Electronic Text Encoding and Interchange*" der *Text Encoding Initiative* (TEI). Ausgehend von einem Überblick über die gegenwärtige Korpuslandschaft wird gezeigt, dass sich in der Kommunikation im Netz trotz des beständigen technologischen Wandels stabile Interaktionsformate etabliert haben. Eine standardisierte Repräsentation solcher Formate bildet eine wichtige Voraussetzung für die Herstellung einer sprachen- wie domänenübergreifenden Interoperabilität von Korpora und leistet einen Beitrag zum Aufbau der Sprachressourcen-Infrastruktur der Zukunft.

Keywords: Digital Humanities, Interaktion, Internetbasierte Kommunikation, Sprachkorpora, TEI

> *„Das Internet? Gibt's diesen Blödsinn immer noch?"*
> (Homer Simpson)

1 Einleitung

Seit 2013 befasst sich eine Arbeitsgruppe (*special interest group*) im Rahmen der *Text Encoding Initiative* (TEI) unter dem Titel *Computer-mediated communication* (kurz: TEI-CMC-SIG) mit der Entwicklung eines XML-Schemas für die Repräsentation und Strukturannotation von Sprachdaten aus Formen internetbasierter Kommunikation. Das von der Gruppe entwickelte Schema soll eine einheitliche und softwareunabhängige, texttechnologische Modellierung von Sprachdaten internetbasierter Kommunikation in linguistischen Korpora ermöglichen und sich zugleich als Austauschformat für Korpusdaten eignen. Das Schema ist als

Michael Beißwenger, Universität Duisburg-Essen, Institut für Germanistik, Berliner Platz 6–8, D-45127 Essen, E-Mail: michael.beisswenger@uni-due.de

https://doi.org/10.1515/9783110538663-015

ein *Basis*schema konzipiert, dessen Grundstruktur sich aus Strukturinformationen, die in den Korpus-Ausgangsdaten (die beispielsweise im HTML-Format vorliegen) bereits enthalten sind, durch einfache Transformationen möglichst automatisch erzeugen lassen soll. Die Motivation für die Erarbeitung eines solchen Basisschemas zielt auf die Bereitstellung einer Lösung für die Aufgabe der Strukturannotation, die für viele Korpusprojekte (zu unterschiedlichen Sprachen und für Daten aus unterschiedlichen IBK-Formen) praktikabel ist und die sich projektspezifisch erweitern lässt. Die Entscheidung, ein solches Schema auf der Grundlage des Encoding-Standards der TEI zu entwickeln, ist motiviert durch dessen Flexibilität: TEI-Schemas lassen sich über den Mechanismus der *customization* erweitern und an individuelle Bedürfnisse anpassen.[1] Daneben ist die Entscheidung für TEI aber auch dadurch motiviert, dass die von der TEI bereitgestellten Formate im Bereich der Digital Humanities als ein *De-facto*-Standard etabliert sind: Viele Sprachressourcen und Korpora – zum Beispiel das Deutsche Referenzkorpus DeReKo am Institut für Deutsche Sprache (IDS) und die DWDS-Textkorpora an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) – nutzen diese Formate bereits (Lüngen & Sperberg-McQueen 2012, CLARIN-D User Guide 2012: Kap. 6). Wer sich dafür entscheidet, seine Sprachressourcen in TEI zu repräsentieren, der sichert seinen Ressourcen damit eine grundsätzliche *Interoperabilität* mit anderen Sprachressourcen: Repräsentationsformate, die von vielen verwendet werden und die darüber hinaus – wie im Falle der TEI-Formate – softwareunabhängig sind, werden mit sehr hoher Wahrscheinlichkeit auch in 20 Jahren noch gepflegt und von gängigen korpustechnologischen Werkzeugen verarbeitet werden können (*Nachhaltigkeit*). Interoperabilität vereinfacht darüber hinaus aber auch die Vernetzung von Ressourcen unterschiedlicher Urheber (*Kombinierbarkeit*). Ressourcen, die auf zumindest basaler Ebene in einem einheitlichen XML-Format repräsentiert sind, lassen sich mit weniger technischem und konzeptionellem Aufwand zusammenführen und mit den gleichen Werkzeugen vergleichend auswerten als Ressourcen, die völlig unterschiedliche Formate verwenden. Nicht zuletzt verringert der Rückgriff auf Standards den Aufwand beim Aufbau neuer Ressourcen: Standards dienen dazu, „dass nicht jeder das Rad neu erfinden muss“ (Lobin 2010: 107); sie ermöglichen, dass für die zu lösende Aufgabe (im hier besprochenen Fall die Aufgabe der Repräsentation und Strukturannotation von Korpora internetbasierter Kommunikation) auf Lösungen zurückgegriffen kann, die sich in anderen Projekten ähnlicher

1 Siehe hierzu das Kapitel *Customization* in den TEI-Guidelines: http://www.tei-c.org/Guidelines/Customization/index.xml (letzter Zugriff: 8. 11. 2017).

Art bereits bewährt haben (*Ökonomie* und *Praktikabilität*). Die Idee der Standardisierung zielt darauf, das Gemeinsame zu erfassen und einheitlich zu beschreiben, das für einen Gegenstand bzw. eine Domäne übergreifend zu den spezifischen Anforderungen und Forschungsinteressen in einzelnen Projektzusammenhängen festgestellt werden kann. Standardisierung und das, was ein Standard abbilden kann (und sollte), hat damit notwendigerweise Grenzen (Perkuhn, Keibel & Kupietz 2012: 68). Gerade deshalb zielt die Arbeit der TEI-CMC-SIG auf die Entwicklung lediglich eines Basisschemas, das grundlegende Struktureigenschaften erfasst, deren Beschreibung in vielen Projekten von Interesse ist. Auf dieses Basisschema sollen weitere, projektspezifische Annotationen aufsetzen können.

In ihrer gegenwärtigen Fassung P5 enthalten die TEI-Guidelines noch keine Modelle für die Strukturbeschreibung von Formen internetbasierter Kommunikation. Die Entwicklung von Lösungen erfolgt daher gegenwärtig auf der Ebene von *customizations*. *Customization* bezeichnet eine Strategie bei der Entwicklung von TEI-Schemas, die es erlaubt, TEI in Domänen anzuwenden, die der Standard in seiner aktuellen Version noch nicht erfasst:

> Because the TEI Guidelines must cover such a broad domain and user community, it is essential that they be customizable: both to permit the creation of manageable subsets that serve particular purposes, and also to permit usage in areas that the TEI has not yet envisioned. Customization is a central aspect of TEI usage and the Guidelines are designed with customization in mind. (TEI-P5: Customization; Hervorhebung MB)

Bislang liegen drei *customizations* für die Strukturannotation von IBK-Korpora vor, die im Zusammenhang mit der Arbeit der TEI-CMC-SIG entwickelt und die bereits in verschiedenen Korpusprojekten zum Deutschen und zum Französischen eingesetzt werden. Nächste Schritte der Arbeit der TEI-CMC-SIG werden darin bestehen, ausgehend von den entwickelten *customizations* eine Eingabe in den Standardisierungsprozess der TEI zu formulieren, um Modelle für die Strukturannotation von IBK-Korpora als echte Elemente einer künftigen Version des TEI-Standards zu verankern.

Der vorliegende Beitrag beschreibt die Zielsetzung und den Ausgangspunkt sowie den aktuellen Stand der Schemaentwicklung der TEI-CMC-SIG. Dazu wird zunächst ein Überblick über die Abdeckung internetbasierter Kommunikation in der gegenwärtigen Korpuslandschaft und über Desiderate im Zusammenhang mit dem Aufbau von Korpora internetbasierter Kommunikation gegeben (Abschnitt 2). Anschließend wird die Frage diskutiert, inwieweit die internetbasierte Kommunikation, die durch beständigen technologischen Wandel geprägt ist, überhaupt sinnvoll in einem Repräsentationsformat abgebildet werden kann, das beansprucht, stabile *Interaktions*formate abzubilden

(Abschnitt 3). Abschnitt 4 gibt einen Überblick über den in 2017 erreichten Stand der Schemaentwicklung. Der Beitrag schließt mit einem Ausblick auf künftige Arbeiten und auf die Korpuslandschaft von morgen (Abschnitt 5).

2 Internetbasierte Kommunikation und Korpuslinguistik: Lagebericht, Forschungsbaustellen und Projekte

Die wissenschaftliche Beschäftigung mit der Kommunikation im Internet kann auf eine gut 25-jährige Geschichte zurückblicken. Einen Fokus der linguistischen Forschung zum Gegenstand bildet dabei von Anfang an die Spezifik dialogisch-interaktionaler Sprachverwendung unter den Bedingungen digitaler Vermittlung. Den so charakterisierten Forschungsgegenstand bezeichne ich als *Internetbasierte Kommunikation* (kurz: IBK).[2] Internetbasierte Kommunikation vollzieht sich in einer Vielzahl unterschiedlicher Kommunikations*formen*, die auf der Verwendung von unterschiedlichen Kommunikations*technologien* beruhen, die je spezifische mediale und modale Ressourcen sowie Bedingungen für die Produktion und Rezeption von kommunikativen Äußerungen vorgeben.[3] Dazu zählen einerseits Technologien, die für die Nutzung anhand von Browsersoftware oder von Clientanwendungen auf dem PC konzipiert sind, und andererseits Technologien, die über sogenannte Apps auf Smartphones und Tabletcomputern als Endgeräten bereitgestellt werden. Eine Kommunikations-

2 Zum Gegenstandsbereich gibt es in der Forschungsliteratur unterschiedliche Vorschläge zur terminologischen Konzeptualisierung, die letztlich einen hohen Grad an extensionaler Übereinstimmung aufweisen. Am ältesten und verbreitetsten ist die Etikettierung als *computer-mediated communication* (CMC, z. B. Herring 1996), ins Deutsche lehnübersetzt als *Computervermittelte Kommunikation*. Der hier präferierte Terminus *Internetbasierte Kommunikation* (IBK), der um die Jahrtausendwende als zeitgemäßere Alternative zu CMC geprägt und u. a. im DFG-Netzwerk *Empirikom* (Beißwenger 2017) verwendet wurde, grenzt die Kommunikation auf Basis von TCP/IP von anderen Formen computervermittelter Kommunikation ab (auch Briefe und Telefongespräche werden heutzutage unter Beteiligung von Computern vermittelt). Jucker & Dürscheid (2012) schlagen die Bezeichnung *Keyboard-to-screen-Kommunikation* vor, die die Spezifik der Ein-/Ausgabedimension fokussiert. *Cum grano salis* werden unter allen diesen Etiketten Erscheinungsformen interaktionaler Sprache unter den Bedingungen digitaler Vermittlung auf Basis von Computernetzwerken untersucht. Für eine eingehendere Diskussion der verschiedenen Konzeptualisierungen sei auf Storrer (2018) verwiesen.

3 Zur Unterscheidung von Kommunikationstechnologien und Kommunikationsformen vgl. Beißwenger (2007: 107–112).

technologie kann dabei exklusiv für den Betrieb einer einzigen Anwendung konzipiert sein, deren Anbieter zugleich der Entwickler[4] der Technologie ist; Beispiele hierfür sind die Technologien, auf denen Social-Network-Plattformen wie Facebook, Google+, Twitter oder Instagram, die Video-Plattform YouTube, Lernplattformen wie moodle und stud.IP oder Kommunikationsdienste wie WhatsApp, Threema, Snapchat oder Skype beruhen. Andere Kommunikationstechnologien existieren in Form von Software, die für die Installation und den Betrieb durch potenziell beliebig viele Anbieter bereitgestellt wird. Dazu zähle ich u. a. E-Mail-, Foren-, Chat-, Weblog- und Wiki-Software, die kostenfrei oder gegen Entgelt für eigene Instanziierungen entsprechender Anwendungen genutzt werden kann. Kommunikation zwischen Nutzern zu ermöglichen, kann die primäre Funktion einer Kommunikationstechnologie sein (Chats, Foren, WhatsApp, Skype) oder auch nur eine Funktion neben weiteren darstellen (wie z. B. im Falle von sozialen Netzwerken, Lernplattformen, Wikis und YouTube). Innerhalb von Kommunikations*formen* sind verschiedene kommunikative *Gattungen* realisierbar – vgl. hierzu analog die Unterscheidung der Konzepte „Kommunikationsform" und „Textsorte" bei Brinker, Cölfen & Pappert (2014: 140–142) –, speziell in Bezug auf die internetbasierte Kommunikation entstehen dabei auch neue Gattungen, die zwar Vorläufer, aber keine direkten Entsprechungen im Bereich der gesprochenen Sprache bzw. der nicht-internetbasierten Kommunikation haben. Kandidaten für solche neuen Gattungen sind in Chats z. B. die „interaktiven Lesespiele" (Beißwenger & Storrer 2012), in Online-Computerspielen die multimodalen „Raids" (Stertkamp 2016), in sozialen Netzwerken die sogenannten „Cybermobbing-Diskurse" (Marx 2015).[5]

Seit ihren Anfängen um Anfang der 1990er Jahre ist die IBK-Forschung stark empirisch ausgerichtet. Die starke empirische Ausrichtung steht aber – bis heute – in einem deutlichen Missverhältnis zur Verfügbarkeit frei zugänglicher Korpora, die für die linguistische Recherche und Analyse aufbereitet sind und die – wie das im Bereich der Textkorpora (z. B. Lüngen 2017; Geyken et al. 2017) bereits möglich ist – als Referenzressourcen für unterschiedlichste Arten von Forschungsfragen genutzt werden können. In aller Regel muss, wer ein sprachliches oder interaktionales Phänomen in der internetbasierten Kommunikation empirisch untersuchen möchte, ein geeignetes Datenset für sein Vorhaben selbst aufbauen. Das kostet Zeit, konzeptionelle Arbeit und häufig auch

4 Aus Gründen der besseren Lesbarkeit wird in diesem Beitrag für alle Personenbezeichnungen das generische Maskulinum gewählt. Die weibliche Form wird dabei stets mitgedacht. Sämtliche Personenbezeichnungen schließen somit beiderlei Geschlecht ein.

5 Zur Abgrenzung von Kommunikationsformen und Gattungen in Bezug auf die internetbasierte Kommunikation vgl. auch Dürscheid (2005).

Geld – Ressourcen, die sich gewinnbringender einsetzen ließen, wenn frei nutzbare Referenzkorpora zum Gegenstand existierten, die verschiedene Kommunikationsformen und -kontexte abdecken und somit zwar sicherlich nicht für alle, aber zumindest für eine große Zahl von Forschungsfragen als Datengrundlage herangezogen werden können.

Die unzureichende Abdeckung der internetbasierten Kommunikation in der aktuellen Korpuslandschaft ist nicht nur für die IBK-Forschung im engeren Sinne ein Problem, sondern generell für jedwede Forschung und praktische linguistische Tätigkeit, die auf die Beschreibung der deutschen Gegenwartssprache zielt: Angesichts der Bedeutung internetbasierter Kommunikation in vielen Bereichen des beruflichen und privaten Alltags und mit Blick auf die schiere Menge an geschriebener und gesprochener Sprache, die tagtäglich unter Nutzung internetbasierter Kommunikationstechnologien produziert wird, ist der Gegenstand „Deutsche Gegenwartssprache" in gegenwartssprachlich ausgerichteten Korpora und Korpussammlungen nur unzureichend erfasst, solange die Sprachverwendung in der internetbasierten Kommunikation darin nur unzureichend abgedeckt ist. Das schließt auch die lexikographische und grammatikographische Beschreibung der deutschen Gegenwartssprache und ihrer Varianten ein.

Um die Abdeckung der internetbasierten Kommunikation in der Korpuslandschaft zum Deutschen und zu anderen Sprachen zu verbessern, müssen die Korpuslinguistik und ihre Nachbardisziplinen (insbesondere die Sprach- und Texttechnologie) sich verschiedener Desiderate und Problembereiche annehmen, die sich im Zusammenhang mit der Erhebung, der Dokumentation, der Aufbereitung und der Wiederbereitstellung von IBK-Daten in Korpora stellen. Diese Desiderate sind schon länger bekannt (vgl. die Problemaufrisse in Beißwenger & Storrer 2008; Storrer 2014: 187–191; Bolander & Locher 2014; Lemnitzer & Zinsmeister 2015: 151–152). Entwicklungs- und Klärungsbedarf besteht insbesondere in Bezug auf die folgenden Fragenkomplexe:

- Fragen der Bewertung des rechtlichen Status für die Erhebung und Bereitstellung von IBK-Daten (exemplarisch iRights.Law Rechtsanwälte 2016; Beißwenger et al. 2017b), damit verbunden die Entwicklung von Konzepten und Verfahren für die Anonymisierung und Pseudonymisierung von IBK-Daten (hierzu u. a. DiDi 2015; Beißwenger et al. 2017b).
- Fragen der Erhebung von IBK-Daten aus Domänen privater Kommunikation (Facebook, WhatsApp, SMS …) (Dürscheid & Stark 2011; Frey, Stemle & Glaznieks 2014; Verheijen & Stoop 2016).
- Entwicklung von Formaten für die Repräsentation von IBK-Daten – als Voraussetzung für den Austausch und die Vernetzung von Ressourcen (Beißwenger et al. 2012; Chanier et al. 2014).

- Entwicklung von Formaten für die Erfassung und Repräsentation von Metadaten; hier sind u. a. auch solche Metadaten zu berücksichtigen, die die Kommunikationsumgebungen beschreiben, aus denen die Korpusdaten erhoben wurden. Mit Blick auf die Veränderlichkeit von digitalen Technologien und Kommunikationsumgebungen sind für die Nachnutzbarkeit der Korpusdaten und für das Verständnis der mit diesen Daten dokumentierten Interaktionen 10, 20 oder 30, unter Umständen sogar bereits 3 Jahre nach ihrer Erhebung solche beschreibenden Daten von großer Wichtigkeit.
- Fragen der Anpassung sprachtechnologischer Verfahren für die Segmentierung und Klassifikation von Sprachdaten (Tokenisierung, morphosyntaktische Annotation, Lemmatisierung, Parsing) an die sprachlichen Besonderheiten internetbasierter Kommunikation (u. a. Giesbrecht & Evert 2009; Horbach et al. 2014; Horsmann & Zesch 2015; WAC-X & EmpiriST 2016).
- Anpassung von Korpusverwaltungs- und -abfragesystemen an die Anforderungen und Nutzerinteressen in Bezug auf IBK-Daten.

Die Gelegenheit für die Erarbeitung von projektübergreifend nützlichen Lösungen für die skizzierten Desiderate – und damit für die Entwicklung von Standards *bottom-up* – ist aktuell günstiger denn je: In den vergangenen Jahren wurden für verschiedene Sprachen und IBK-Formen Korpusprojekte auf den Weg gebracht, deren Ergebnisse nach Abschluss der Projektlaufzeit der Scientific Community als Ressourcen zur Verfügung gestellt werden sollen oder bereits zur Verfügung stehen. Die Erfahrungen und Lösungsansätze, die in diesen Projekten entwickelt werden, stellen wertvolle Ressourcen für die Entwicklung von projektübergreifenden Lösungen dar.[6]

Beispiele für aktuelle (abgeschlossene und laufende) Korpusprojekte sind:

- *CoMeRe*: eine Sammlung von vierzehn IBK-Korpora zum Französischen, die neun verschiedene IBK-Genres abdeckt (SMS, Wikipedia-Diskussionen, Tweets, Weblogs, E-Mails, Diskussionsforen, Chat, multimodale Formen) und die neben rein schriftlichen auch multimodale, synchrone und asynchrone Formen sowie Kommunikation aus dem öffentlichen und aus dem privaten Bereich erfasst. Sämtliche Korpora sind in TEI repräsentiert (Chanier et al. 2014) und werden unter einer CC-BY-Lizenz als OpenData über ORTOLANG[7] als downloadbare Ressourcen zur Verfügung gestellt.

6 Beispiele für Lösungsansätze aus verschiedenen Projekten beschreiben Beißwenger et al. (2017a).

7 http://hdl.handle.net/11403/comere (letzter Zugriff: 8. 11. 2017).

- *CorCenCC-CMC*: eine seit 2016 im Aufbau befindliche *e-language*-Komponente im Projekt *Corpws Cenedlaethol Cymraeg Cyfoes* (*National Corpus of Contemporary Welsh*, CorCenCC) mit Sitz an der Cardiff University (UK).[8]
- *DeReKo-News*: das seit 2013 aufgebaute deutsche Newsgroup-Korpus in DeReKo[9] im Umfang von 98 Millionen Tokens mit Daten aus den Jahren 2013–2016 (Schröck & Lüngen 2015).
- *DeReKo-Wikipedia*: die Wikipedia-Korpora in DeReKo, die Artikel- und Diskussionsseiten im Umfang von 581 Millionen Tokens enthalten und vom IDS sowohl als downloadbare Ressourcen als auch zur Abfrage über COSMAS II[10] angeboten werden (Margaretha & Lüngen 2014).
- *DiDi*: Korpus zur Sprachverwendung in Facebook mit Daten deutscher und italienischer Sprache sowie in Südtiroler Mundart, aufgebaut im Rahmen des DiDi-Projekts (*Digital Natives – Digital Immigrants*) an der EURAC in Bozen (IT) und online abfragbar via ANNIS (Frey, Glaznieks & Stemle 2016).[11]
- *Dortmunder Chat-Korpus*: Korpus mit einer Million laufenden Wortformen zur deutschsprachigen Chat-Kommunikation in unterschiedlichen Handlungsbereichen (Freizeit, Bildung, Beratung, Medien), die um XML-Annotationen zu ausgewählten Sprachmerkmalen angereichert wurden (Beißwenger 2013). Das Korpus wurde seit 2005 in einer „Releaseversion" zum freien Download angeboten.[12] Eine um *Part-of-speech*-Annotationen sowie eine TEI-Repräsentation erweiterte, vollständig anonymisierte Version (*Chat-Korpus 2.0*) wurde 2015–2017 in die CLARIN-D-Korpusinfrastrukturen integriert und steht seit Herbst 2017 als Teil des *DeReKo* und der Korpussammlung der Berlin-Brandenburgischen Akademie der Wissenschaften für die Online-Abfrage über die Korpus-Rechercheschnittstellen am Institut für Deutsche Sprache (IDS) Mannheim und über das Portal www.dwds.de zur Verfügung (Lüngen et al. 2016; Beißwenger et al. 2017b).[13]

8 http://sites.cardiff.ac.uk/corcencc/ (letzter Zugriff: 8. 11. 2017).

9 http://www.ids-mannheim.de/dereko (letzter Zugriff: 8. 11. 2017).

10 https://cosmas2.ids-mannheim.de/ (letzter Zugriff: 8. 11. 2017).

11 http://www.eurac.edu/didi (letzter Zugriff: 8. 11. 2017).

12 http://www.chatkorpus.tu-dortmund.de/ (letzter Zugriff: 8. 11. 2017).

13 Am IDS ist das Chat-Korpus über COSMAS II recherchierbar und darüber hinaus via http://hdl.handle.net/10932/00-0379-FDFE-CC30-0301-E (letzter Zugriff: 8. 11. 2017) als downloadbare Ressource über das IDS-Repository erhältlich. Im BBAW-Repository steht das Korpus unter http://hdl.handle.net/11858/00-203Z-0000-002D-EC85-5 (letzter Zugriff: 8. 11. 2017) zur Verfügung. Nach kostenfreier Registrierung können Interessierte das Korpus im Portal www.dwds.de im Bereich „Textkorpora" abfragen.

- *DWDS-Blogkorpus*: 103 Millionen Tokens aus CC-lizenzierten Weblogs, großenteils in deutscher Sprache, die über das DWDS-Portal[14] der BBAW recherchierbar sind (Barbaresi & Würzner 2014; Barbaresi 2016).
- *Gießener Scienceblog-Korpus*: laufendes Projekt zum Aufbau eines Korpus deutschsprachiger Wissenschaftler-Blogs an der Universität Gießen (Grumt Suárez, Karlova-Bourbonus & Lobin 2016).
- *Janes*: Projekt *Jezikoslovna analiza nestandardne slovenščine* (*Corpus of Nonstandard Slovene*, Fišer, Erjavec & Ljubešić 2016, 2017)[15] mit 200 Millionen Tokens aus verschiedenen IBK-Formen (Tweets, Forendiskussionen, Blogs, Leserkommentare aus Nachrichtenportalen, Wikipedia-Diskussionsseiten). Die Korpusdaten sind linguistisch annotiert und wurden automatisch nach dem Grad ihrer Konformität zum geschriebenen Standard klassifiziert (Ljubešić et al. 2015); die dabei ermittelten Werte sind den Daten in Form von Metadaten beigegeben.
- *MoCoDa²*: laufendes Projekt an der Universität Duisburg-Essen, in dem eine Datenbank und ein Web-Frontend für die wiederholte Sammlung von Spenden aus digitaler Kurznachrichtenkommunikation (WhatsApp und vergleichbare Dienste) entwickelt wird.[16] Die Datenstücke der Sammlung werden unter Einbeziehung der Spender mit reichhaltigen Metadaten ausgestattet und dadurch insbesondere auch für qualitative korpusgestützte Analysen interessant. Es ist geplant, die aufbereiteten Datenspenden in regelmäßigen Abständen in die Korpussammlungen am IDS zu integrieren.
- *NPS Chat Corpus*: ein Korpus mit 45.000 laufenden Wortformen aus englischsprachigen, altersspezifischen Chatrooms, das um *Part-of-speech*-Informationen und eine Dialogaktklassifikation angereichert ist (Forsyth & Martell 2007) und das über das *Linguistic Data Consortium* (LDC) zur Verfügung gestellt wird.[17]
- *sms4science.ch*: Korpus mit gespendeten SMS-Nachrichten (Deutsch, Französisch, Schweizerdeutsch, Italienisch, Rätoromanisch) im Umfang von 650.000 Tokens (Dürscheid & Stark 2011); online abfragbar in einer Volltextversion (SMS Navigator) und als teilweise annotierte Version in ANNIS.[18]
- *SoNaR-CMC*: IBK-Komponente mit Chats, Tweets und SMS-Nachrichten im Referenzkorpus des Gegenwarts-Niederländischen (Oostdijk et al. 2013); online abfragbar via CLARIN-NL (OpenSoNaR).[19]

14 https://www.dwds.de (letzter Zugriff: 8. 11. 2017).
15 http://nl.ijs.si/janes/ (letzter Zugriff: 8. 11. 2017).
16 http://www.mocoda2.de/ (letzter Zugriff: 8. 11. 2017).
17 http://faculty.nps.edu/cmartell/NPSChat.htm (letzter Zugriff: 8. 11. 2017).
18 http://www.sms4science.ch (letzter Zugriff: 8. 11. 2017).
19 https://portal.clarin.nl/node/4195 (letzter Zugriff: 8. 11. 2017).

- *Suomi24*: umfangreiche Sammlung (2,38 Milliarden Tokens) mit Daten aus finnischen Diskussionsforen mit morphosyntaktischen Annotationen; als Download verfügbar.[20]
- *Web2Corpus_it*: laufendes Projekt zum Aufbau eines ausgewogenen Korpus zur italienischen IBK mit Daten aus Foren, Blogs, Newsgroups, sozialen Netzwerken und Chats (Chiari & Calzonetti 2014).[21]
- *whatsup-switzerland.ch*: Korpus des Projekts *What's up, Switzerland?*, in dem spendenbasiert 650 WhatsApp-Chatverläufe im Umfang von insgesamt 5 Millionen Tokens gesammelt wurden.[22]

Der Austausch von *best practices*, Werkzeugen und Expertise zwischen Korpusprojekten zum Thema stellt eine wichtige Vorbedingung dar, um in einem *community-driven approach* projektübergreifend nutzbare, dokumentierte und nachhaltige Lösungen für die Herausforderung zu entwickeln, internetbasierte Kommunikation in linguistischen Korpora zu repräsentieren. Als fruchtbares Format, um die dafür erforderliche Vernetzung anzustoßen, hat sich die Konferenzreihe „Conference on CMC and Social Media Corpora in the Humanities" erwiesen, bei der seit 2013 in jährlichem Turnus Korpusprojekte, Entwickler von korpus- und sprachtechnologischen Werkzeugen sowie korpuslinguistisch arbeitende Geisteswissenschaftler zusammenkommen, um aktuelle Forschungs- und Entwicklungsarbeiten für verschiedene Sprachen vorzustellen sowie Erfahrungen und Lösungsansätze zu diskutieren.[23] Die enge Anbindung an Initiativen und Verbundprojekte, die sich im Bereich der Digital Humanities mit der Etablierung von einheitlichen Formaten (TEI) und Sprachressourceninfrastrukturen (CLARIN, DARIAH) beschäftigen, stellt die Voraussetzung dar, um die Interoperabilität von IBK-Korpora mit Sprachressourcen zu anderen Domänen des Sprachgebrauchs (Text- und Gesprächskorpora) zu gewährleisten.

Interoperabilität: Korpora, die in aufeinander abbildbaren Formaten repräsentiert sind, können einfacher miteinander kombiniert und mit den gleichen Korpustechnologien ausgewertet werden als wenn das nicht der Fall ist. Sind diese Formate zudem kompatibel mit Formaten, die in existierenden Text- und

20 http://urn.fi/urn:nbn:fi:lb-201412171 (letzter Zugriff: 8. 11. 2017).

21 Project page: http://www.glottoweb.org/web2corpus/ (letzter Zugriff: 8. 11. 2017).

22 http://www.whatsup-switzerland.ch/ (letzter Zugriff: 8. 11. 2017). Eine vergleichbare, einmalige Datensammlung für das Deutsche (http://www.whatsup-deutschland.de/, letzter Zugriff: 8. 11. 2017) wurde 2014/2015 von Beat Siebenhaar (Leipzig) initiiert und unter Beteiligung von sieben deutschen Universitätsstandorten durchgeführt.

23 Einen Überblick über Korpusprojekte und darauf bezogene Forschungsfragen bieten die Buchpublikationen zu den Konferenzen 2013, 2015 und 2016 in Dortmund, Rennes und Ljubljana (Beißwenger et al. 2014; Wigham & Ledegen 2017; Fišer & Beißwenger 2017).

Gesprächskorpora verwendet sind, können IBK-Korpora darüber hinaus auch mit Korpora anderen Typs kombiniert und in existierende Korpussammlungen integriert werden. Für die empirische Analyse von Sprache und Interaktion eröffnet das interessante Perspektiven:

1. verbesserte Möglichkeiten der kombinierten Auswertung von IBK-Korpora in unterschiedlichen Sprachen im Rahmen sprachvergleichender Untersuchungen;
2. verbesserte Möglichkeiten der Untersuchung von sprachlichen und kommunikativen Praktiken in unterschiedlichen Formen und Gattungen internetbasierter Kommunikation, die in Korpora unterschiedlicher Anbieter dokumentiert sind;
3. verbesserte Möglichkeiten der kombinierten Auswertung von IBK-Korpora mit Text- und Gesprächskorpora (vergleichende Untersuchung sprachlicher und interaktionaler Phänomene in monologischen Texten, mündlichen Gesprächen und in dialogischer Schriftlichkeit).

Ein wichtiger Baustein für die Herstellung von Interoperabilität ist die Verwendung eines einheitlichen Basisformats für die Repräsentation und Strukturbeschreibung der Korpusdaten. Ein solches Format existiert bislang nicht, es wird aber dringend benötigt.

3 Panta rhei? Internetbasierte Kommunikation zwischen Veränderlichkeit und Stabilität

Die Entwicklung eines Basisformats für die Repräsentation von Daten internetbasierter Kommunikation setzt voraus, dass es übergreifende und stabile Interaktionsformate gibt, die sinnvoll in einem einheitlichen Repräsentationsformat abgebildet werden können. Ein Repräsentationsformat, das für möglichst viele Korpora als grundlegendes Annotations- und als Austauschformat brauchbar sein soll, darf weder so spezifisch sein, dass für jedes Genre internetbasierter Kommunikation (z. B. Chat, SMS, Instant Messaging, Microblogging, Forum, Blogkommentare) bzw. für Daten aus unterschiedlichen Plattformen und Anwendungen (z. B. WhatsApp, Threema, Facebook, Instagram, Twitter, Wikipedia-Diskussionen) ein eigenes Strukturmodell benötigt wird. Stattdessen sollte es einen Abstraktionsgrad aufweisen, der zentrale Strukturmerkmale internetbasierter Kommunikation übergreifend zu Kommunikationsformen und -anwendungen erfasst. Das mit ihm beschriebene Interaktionsformat sollte zudem so stabil sein, dass mit hoher Wahrscheinlichkeit davon ausgegangen

werden kann, dass das Strukturmodell nicht Jahr für Jahr einer Revision bedarf, um es an zwischenzeitlich eingetretene Veränderungen der Technologie anzupassen. Ein Repräsentationsformat, das sich ständig verändert, gewährleistet gerade keine Interoperabilität.

Nun ist jedoch insbesondere die internetbasierte Kommunikation in hohem Maße von Veränderlichkeit geprägt. Das Nachdenken über ein Strukturmodell muss daher zunächst die folgenden Fragen klären:

- Gibt es überhaupt übergreifende Interaktionsformate in der internetbasierten Kommunikation?
- Wenn ja: Sind diese stabil, d. h. relativ beständig gegenüber technologischem Wandel?

Unter einem *Interaktionsformat* verstehe ich dabei eine Struktur für die Organisation von Interaktionsereignissen, die durch kommunikative Rahmenbedingungen geprägt ist, die sich aus Festlegungen auf der Ebene der Kommunikationstechnologie ergeben und für die Nutzer der Technologie nicht änderbar sind. Die Gesamtheit dieser Bedingungen definiert die medialen und modalen Ressourcen, die für die Interaktions*praxis* auf Grundlage der betreffenden Technologie zur Verfügung stehen. Da sich das Format nicht als solches, sondern immer erst bei seiner Instanziierung in konkreten Interaktionsereignissen zeigt, stellt die Konkretisierung des Formats im Vollzug immer die Schnittstelle zwischen den invarianten technologischen Rahmenbedingungen und den sprachlichen und kommunikativen Praktiken dar, mit denen sich die Interagierenden unter Nutzung der gegebenen Ressourcen die technologischen Bedingungen für die Zwecke der Interaktionskonstitution zu eigen machen.

3.1 Veränderlichkeit

Im Kontext der Modellierung und Analyse sprachlicher und kommunikativer Praktiken (Deppermann et al. 2016) bildet die internetbasierte Kommunikation einen spannenden Untersuchungsgegenstand. Geradezu *Fishbowl*-artig lässt sich an (und in) ihr studieren, wie sich die Nutzer internetbasierter Kommunikationstechnologien an die von der Technologie gesetzten Rahmenbedingungen anpassen, um unter diesen Bedingungen bestmöglich das zu tun, wozu die Technologie gemacht ist: Interaktion zu gestalten. Da jedweder Austausch im Netz – das schließt browserbasierte Anwendungen und auf dem Smartphone genutzte, App-basierte Anwendungen gleichermaßen ein – nur unter den Bedingungen technischer Vermittlung zustande kommt und die technischen Rahmenbedingungen für die Interaktionsorganisation in jeder Anwendung unterschiedlich ausgeprägt sein können, bildet die Analyse internetbasierter

Kommunikation ein äußerst produktives Anwendungsfeld, um die Vielfalt und Flexibilität der Praktiken-Praxis unter Bedingungen technischer Vermittlung zu untersuchen. Nicht von ungefähr nimmt die Analyse von Praktiken breiten Raum in der interaktions- und konversationsanalytischen Beschäftigung mit internetbasierter Kommunikation ein. Stark beforscht wurden in der IBK-Forschung schon früh beispielsweise Ausprägungen der Konzepte *turn*, *turn-taking*, *adjacency* und *floor* (Murray 1989; Garcia & Baker Jacobs 1998, 1999; Cherny 1999; Storrer 2001; Beißwenger 2003 u. a.), Praktiken der Kohärenzsicherung (Herring 1999; Severinson Eklundh 2010), die Beitragsproduktion bzw. Turnkonstruktion (Garcia & Baker Jacobs 1998, 1999; Markman 2006; Beißwenger 2007), Formen der Reparatur und der schriftlichen Revision (Schönfeldt & Golato 2003; Beißwenger 2010). Neuere Arbeiten, die Phänomene internetbasierter Kommunikation auf dem Hintergrund des Programms der Interaktionalen Linguistik (Selting & Couper-Kuhlen 2000) analysieren, sind z. B. Imo (2015a) zur konstruktionsgrammatischen Analyse von Emoticons, Imo (2015b) zur Portionierung von Äußerungen in WhatsApp-Dialogen, König (2015) zu Sequenzmustern, Lindemann, Ruoss & Weinzinger (2014) zu Praktiken des Editierens. Einen Überblick über Kandidaten für Praktiken in verschiedenen Formen internetbasierter Kommunikation bietet Beißwenger (2016).

Praktiken variieren in Abhängigkeit zu den modalen und medialen Ressourcen, die den Interaktionsbeteiligten für die Interaktionskonstitution zur Verfügung stehen. Die Ressourcen, die in Formen der technisch vermittelten Kommunikation genutzt werden können, variieren in Abhängigkeit zur Veränderlichkeit von Technologien. In technologischer Hinsicht ist Veränderlichkeit ein fundamentales Merkmal der internetbasierten Kommunikation: Neue Kommunikationstechnologien kommen auf und verdrängen andere Kommunikationstechnologien aus der Gunst der Nutzer bzw. bestimmter Nutzergruppen. So hat beispielsweise die Anwendung Snapchat nach den Ergebnissen der jüngsten Ausgabe der JIM-Studie (2016) in der Gunst der 12–19-Jährigen deutlich zugelegt, während für Facebook bei derselben Gruppe ein vergleichbar großer Abfall in der Nutzergunst zu verzeichnen ist. Zugleich entwickeln sich Kommunikationstechnologien weiter, so dass sich für die Nutzer der auf ihrer Grundlage bereitgestellten Anwendungen die zur Verfügung stehenden technischen Handlungsmöglichkeiten und damit die medialen und modalen Ressourcen für die Interaktionskonstitution kontinuierlich verändern. So hat die Plattform Facebook 2013 die ursprünglich nur auf Twitter nutzbaren Hashtags eingeführt, um ihren Nutzern zusätzliche Möglichkeiten zur thematischen Vernetzung von Beiträgen zu bieten. Durch diese Änderung hat sich das System von Mitteln für die thematische Organisation und Verknüpfung von Interaktionen insgesamt verändert. Ähnliches gilt für die – ebenfalls von Twitter

übernommene – Einführung der automatischen Verlinkung von Adressierungsausdrücken auf die Profilseite des adressierten Nutzers, verbunden mit der Anzeige des Postings, das eine Adressierung enthält, auf der individuellen Startseite des Adressaten: Adressierung wird dadurch in Facebook zu einem Mittel, um die Aufmerksamkeit von Adressaten zu gewinnen und diesen die Existenz für sie relevanter Interaktionsäußerungen im *Push*-Verfahren bekannt zu machen. Damit gewinnt die Praktik des Adressierens, wie sie sich in der frühen Chat- und Foren-Kommunikation als zunächst rein textuelle Praktik für die Kohärenzsicherung herausgebildet hat, an zusätzlichen Dimensionen und entwickelt sich zu einer Praktik der Aufmerksamkeitsgewinnung.

3.2 Stabilität

Das Internet ist mittlerweile alt genug für Stabilität. Bei all der Variabilität von Ressourcen und darauf bezogenen Praktiken gibt es durchaus Formate, die sich als weitgehend beständig gegenüber Wandel erwiesen haben und die sich, auch wenn sie in unterschiedlichen Instanziierungen unterschiedlich ausgeprägt sein mögen, übergreifend zu einzelnen IBK-Genres und -Plattformen beschreiben lassen. Im Folgenden charakterisiere ich anhand einer Gegenüberstellung der Instanziierungen von Interaktion in sechs verschiedenen Arten von Kommunikationsumgebungen ein Format, das seit inzwischen mehr als einem Vierteljahrhundert in der internetbasierten Kommunikation präsent ist und das sicherlich auch noch in den nächsten zehn Jahren Bestand haben wird. Zwar wandelt auch dieses Format kontinuierlich sein Gesicht und stellt sich in unterschiedlichen Instanziierungen durchaus unterschiedlich dar; im Kern hat es sich jedoch als erstaunlich robust erwiesen.

Zwischen den in den Abb. 14.1–14.6 veranschaulichten Instanziierungen von Interaktion gibt es Gemeinsamkeiten und Unterschiede. Gemeinsam ist ihnen, dass sie aus schriftlichen Einheiten aufgebaut sind, die im Layout klar voneinander abgegrenzt sind und die jeweils einem Urheber zugeordnet sind. Ich nenne diese Einheiten *Postings*, weil das englische Verb ‚to post' und sein ins Deutsche entlehntes Pendant ‚posten' treffend ausdrücken, was diese Einheiten charakterisiert und von den grundlegenden Einheiten mündlich realisierter Interaktionen – den Turns – unterscheidet:

- Sie werden erst *als Ganze* für die anderen Beteiligten wahrnehmbar, und zwar erst dann, nachdem sie durch Ausführung einer expliziten Verschickungsanweisung an den Server übergeben und von diesem weiterübermittelt wurden. Eine inkrementelle Verarbeitung der Äußerung (= des Geposteten) simultan zu ihrer Hervorbringung und damit eine rezipienten-

Abb. 14.1: Vier Postings mit Thread-Struktur auf einer Facebook-Profilseite: Die Postings 3 und 4 erscheinen, da sie von den Verfassern als „Antworten" auf das Vorgänger-Posting gekennzeichnet wurden, eingerückt.

seitige Einwirkung auf den Verbalisierungsplan des Produzenten, wie das für Turns in mündlichen Gesprächen charakteristisch ist, ist ausgeschlossen.

– Die Produktion geht der Wahrnehmung durch andere voraus und ist ein privater Akt, auch wenn manche Kommunikationsanwendungen – z. B. WhatsApp – ihren Nutzern über temporäre Hinweise anzeigen, wenn einer der anderen Beteiligten gerade einen Beitrag eingibt („Nutzer X schreibt ..."). Was während der Beitragseingabe im Eingabeformular auf dem Display des Produzenten entsteht und wie es sich verändert, bleibt den übrigen Beteiligten aber verborgen.
– Solange der eingegebene Beitrag nicht verschickt wurde, hat der schriftliche Entwurf den Charakter der Vorläufigkeit – und zwar auch technisch: Er kann ganz oder in Teilen, beliebig oft und in beliebigem Umfang revidiert werden. Beißwenger (2007, 2010) zeigt empirisch an dokumentierten Produktionsverläufen, dass von dieser Möglichkeit rege Gebrauch gemacht wird. Eine

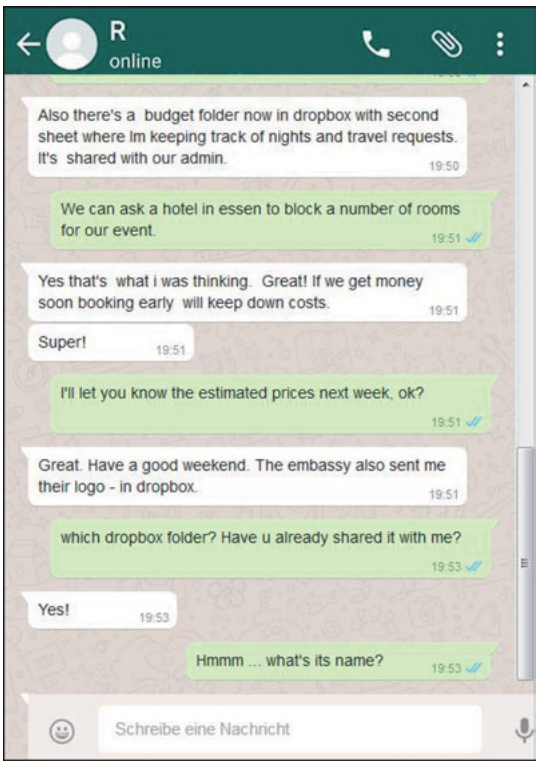

Abb. 14.2: Ausschnitt aus einer WhatsApp-Interaktion. Die rechtsbündig dargestellten Postings wurden von der Interaktionspartnerin verfasst und sind mit automatisch generierten Hinweisen zum Übermittlungsstatus (Häkchen rechts unten) versehen.

> Bearbeitung des Entwurfs ist grundsätzlich nicht nur für den Autor selbst, sondern auch für Zusatzprogramme möglich, die auf dem Endgerät oder in der betreffenden Kommunikationsanwendung im Hintergrund laufen (z. B. Autokorrekturfunktionen, die, insbesondere auf mobilen Endgeräten, z. T. erheblich zur sprachlichen Gestalt der verschickten Postings beitragen).

Einen Sonderfall der Postings stellen die sogenannten Sprachnachrichten in WhatsApp-Interaktionen dar. Dabei handelt es sich um aufgezeichnete, mündlich realisierte Äußerungen, die in Form von Audiodateien übermittelt und dadurch – analog zu ihrem schriftlichen Pendant – erst im Nachhinein zur Verbalisierung rezipierbar werden. Ich bezeichne diese Form der Postings als *Audio-Postings*, um sie – aus denselben Gründen wie im Falle der schriftlichen Postings – terminologisch von Turns in Gesprächen abzugrenzen. Revisionen

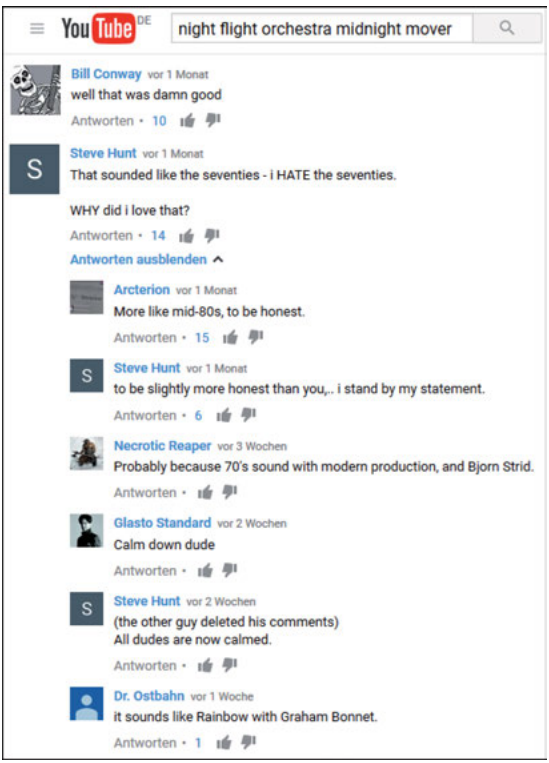

Abb. 14.3: Kommentarsektion unter einem YouTube-Video. Die einzelnen Postings sind – ähnlich wie in Facebook – mit sog. „Likes" versehen; von den Verfassern als Antworten markierte Postings erscheinen eingerückt.

von Sprachnachrichten sind nur komplett möglich: Eine laufende Audioaufzeichnung kann abgebrochen, aber nicht in Teilen editiert werden. Für übermittelte Audio-Postings gilt, dass der Prozess ihrer Verbalisierung in der Rezeption für die Adressaten transparent wird; das gilt aber nicht für Vorversionen, bei denen die Aufnahme abgebrochen wurde. Auch das ist ein wichtiger Unterschied zu den Verbalisierungs- und Wahrnehmungsbedingungen in mündlichen Gesprächen.

Gemeinsam ist den abgebildeten Beispielen weiterhin Folgendes:

– Die Vermittlung der Interaktionsbeiträge erfolgt über ein *Logfile* am Bildschirm, das im Falle „synchroner"[24] Kommunikationsumgebungen (z. B. in

24 Ich verwende den Ausdruck ‚synchron' in Anführungszeichen, um deutlich zu machen, dass Chats aufgrund der oben beschriebenen Zeitlichkeitsbedingungen nicht in gleicher Weise unter Bedingungen der „Gleichzeitigkeit" stattfinden wie mündliche Gespräche. Auf Garcia &

Abb. 14.4: Posting auf Twitter („Tweet") mit zwei Antwort-Postings, die als Thread unter dem Bezugsposting angezeigt werden. Die einzelnen Postings enthalten automatisch generierte Metadaten (u. a. zur Anzahl der Likes und Antworten sowie zur Anzahl der Retweets).

klassischen IRC- und Webchats) mindestens für die Dauer der aktuellen Sitzung, in vielen Umgebungen aber theoretisch unbegrenzt persistent ist (WhatsApp, Online-Foren, Twitter, Facebook usw.). Was geäußert wird, dokumentiert sich in einem gespeicherten Verlauf. Durch das Posting-Format der ausgetauschten Beiträge wird die Persistenz von Interaktionsbeiträgen zur notwendigen Grundbedingung für die Interaktionskonstitution: Wenn Beiträge zunächst als Ganze verbalisiert werden, bevor sie in einem der Verbalisierung nachgeordneten Schritt für die Adressaten wahrnehmbar werden, bedarf die Interaktion der *Überlieferung* (i. S. v. Ehlich 1984), um zu funktionieren. Da überdies der Zeitpunkt, ab dem ein Posting am Bild-

Baker Jacobs (1998) geht der Vorschlag zurück, die Kommunikation in Chats und mit vergleichbaren Technologien als *quasi-synchronous* (bzw. mit Dürscheid 2005 als „quasi-synchron") zu charakterisieren. Das Etikett ‚quasi-synchron' signalisiert zwar, dass die Kommunikation in den damit charakterisierten Form(at)en nicht gänzlich synchron verläuft, bleibt hinsichtlich des Unterschieds, um den es geht, jedoch unterspezifiziert bzw. weist ihn als letztlich doch vernachlässigbar aus (lat. *quasi* = ‚vergleichbar, ungefähr'). Gerade im ‚quasi' zeigen sich allerdings die für die Interaktionskonstitution fundamentalen Unterschiede zum Gespräch. In Beißwenger (2007, 2010) habe ich deshalb vorgeschlagen, Formen wie Chats als „synchrone, aber nicht simultane" Formen bzw. die Kommunikationsbedingungen in Chats und vergleichbaren

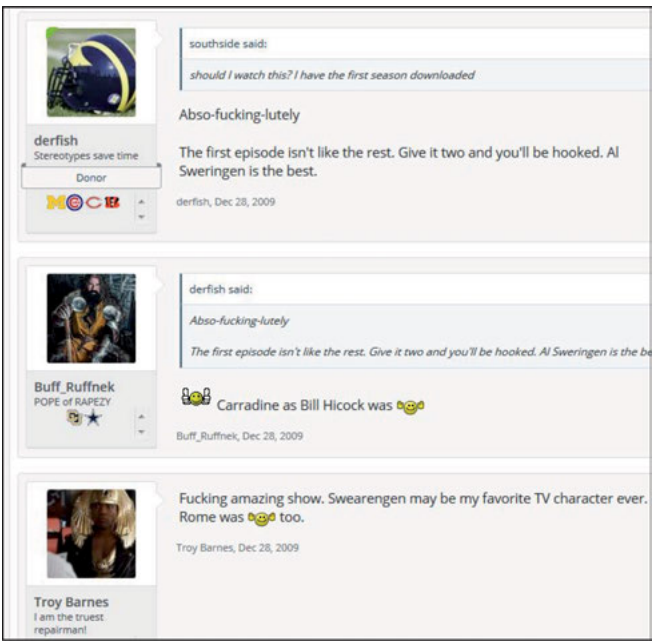

Abb. 14.5: Ausschnitt aus einem Thread mit einem Online-Forum. Die Verfasser der Postings 1 und 2 nutzen die Möglichkeit, vorangegangene Postings anderer Urheber ganz oder in Teilen zu zitieren. Die Zitate werden vom System automatisch in den Beiträgen reproduziert; dabei werden zudem automatisch generierte Sprachbausteine eingefügt („southside said", „derfish said"). Auch die einzelnen Postings nebengestellter Ausschnitte aus den hinterlegten Nutzerprofilen der Verfasser wurden vom System automatisch reproduziert.

schirm wahrnehmbar ist, nicht notwendigerweise mit dem Zeitpunkt seiner tatsächlichen Wahrnehmung durch die Adressaten zusammenfällt, sichert das *Logging* der Postings nicht nur die Bearbeitung der zeitlichen Zerdehnung zwischen Produktion und Zustellung, sondern zudem die Bearbeitung der Zerdehnung zwischen Zustellung und tatsächlicher Rezeption. Dass selbst in „synchronen" Formen wie dem klassischen Chat, in

Formen als „synchron ohne Synchronisierung" zu beschreiben. Synchronizität wird durch diesen Vorschlag in zwei Aspekte zerlegt: (i) den Aspekt der zeitgleichen Orientiertheit der Beteiligten auf die Entwicklung des Kommunikationsgeschehens, die auch der vorwissenschaftlichen und alltagssprachlichen Charakterisierung von Chats als „synchronen" Formen zugrunde liegt; (ii) den Aspekt der Alignierung von Verbalisierung, Übermittlung und Verarbeitung, die im Falle mündlicher Gespräche vollumfänglich gegeben (i. S. v. der von Auer 2000 beschriebenen ‚On-line'-Verbalisierung), im Falle von Chats aber durch eine Zerlegung in eine konsekutive Abfolge von Schritten gekennzeichnet ist (Beißwenger 2007: 35–37; Beißwenger 2010: 257–258).

Offrande au Saint Sacrement [Quelltext bearbeiten]

Dieses Orgelstück kenne ich nicht, da es offensichtlich in keiner Messiaen-Orgel-Gesamtaufnahme enthalten ist. Wie ist das zu erklären? Und welcher Art ist das Stück (groß, klein; mit gregorianischem Thema?) Humpyard 00:34, 28. Jan. 2008 (CET)

> Das weiss ich natürlich auch nicht. Eine Möglichkeit wäre, es handelt sich um Unkenntnis. Vielleicht sind manchmal Titel nicht eindeutig? Vielleicht wird etwas als "Orgel-Gesamtaufnahme" bezeichnet, wenn die Aufnehmer vermuten, es sei gesamt? Letzte Idee: Vielleicht hat es der Komponist nicht ausdrücklich für Orgel ausgewiesen.- All dies sind nur theoretisierende Denkansätze! --888344
>
> > Oder auch - kucke mal hier: http://www.joergabbing.de/publikationen.htm --888344
> >
> > > in der Latry-Gesamtaufnahme (DGG) ist das Stück enthalten (CD2), die Google-Suche fördert außerdem mehrere Quellen zu Tage, die Partitur des Werkes ist z.B. bei Leduc erschienen Gruß Akeuk 20:02, 2. Feb. 2008 (CET)

Abb. 14.6: Ausschnitt aus einer Wikipedia-Diskussionsseite. Der Grad der Einrückung der einzelnen Postings wurde von den Verfassern bei der Eingabe manuell erzeugt.

denen die Interaktionsbeteiligten zeitgleich auf die Teilhabe an der Weiterentwicklung des Interaktionsgeschehens orientiert sind, Beiträge häufig erst im Nachhinein zu ihrer Verfügbarkeit am Bildschirm gelesen werden, zeigt Beißwenger (2007).

- Im Logfile, das als Rezeptionsgrundlage fungiert, erscheinen die Postings in einer klaren Abfolge. Diese muss nicht mit der von den Interaktionsbeteiligten intendierten Handlungsabfolge identisch sein. Die Zerdehnung von Produktion und Übermittlung begünstigt, insbesondere in „synchronen" Gruppeninteraktionen, die Überkreuzung von Posting-Sequenzen, die unterschiedliche Handlungssequenzen realisieren (Herring 1999; Storrer 2001). In „asynchronen" Formen nutzen die Schreiber Möglichkeiten des Threadings und der Arbeit mit Zitaten, um die sequenzielle Einordnung ihrer Beiträge anzuzeigen und nachvollziehbar an bestimmte Positionen der Sequenz anzuknüpfen (z. B. Severinson-Eklundh 2010).
- Das *Logging* des Kommunikationsverlaufs am Bildschirm schafft gegenüber mündlichen Interaktionen veränderte Bedingungen für den Bezug auf frühere Teile der Sequenz: Postings können mehrfach und wiederholt rezipiert werden. Das gilt für medial schriftlich realisierte Postings in gleicher Weise wie für Postings, die in Form aufgezeichneter Audioaufnahmen realisiert sind (Sprachnachrichten in WhatsApp). Wann ein Posting interaktional bearbeitet und beantwortet wird, kann flexibler gehandhabt werden als im mündlichen Gespräch. In Online-Foren, die über viele Jahre

laufen, lässt sich beobachten, dass bisweilen an Postings oder Threads angeknüpft wird, die mehrere Monate oder gar Jahre alt sind. Und selbst in Anwendungen wie WhatsApp, die eine „synchrone" Nutzung erlauben, werden Postings nicht immer unmittelbar beantwortet – auch dann nicht, wenn sie von den Adressaten unmittelbar nach ihrer Zustellung rezipiert wurden. Die Persistenz des Verlaufs erlaubt es vielmehr, die Teilhabe an Interaktionen flexibel an das individuelle Zeit- und Aufgabenmanagement anzupassen.

– Der Interaktionsverlauf ist digital gespeichert und mit Standardwerkzeugen von Betriebssystemen verwalt- und bearbeitbar: Eigene und fremde Postings sowie Teile davon (d. h. auch die darin integrierten Bild-, Ton- und Videodateien) können in die Zwischenablage kopiert und in anderen Kontexten reproduziert werden; die am Bildschirm vorgehaltenen Verläufe können in Dateien gespeichert und weiterverwendet werden.
– Neben dem vom Verfasser eingegebenen sprachlichen Inhalt (*user generated content*) können Postings automatisch, d. h. vom System generierte oder reproduzierte, Beitragsteile enthalten. Dazu zählen beispielsweise die automatisch eingefügten Zitate von Postings oder Posting-Teilen anderer Schreiber in Abbildung 14.5, im weiteren Sinne aber auch die vom System hinzugefügten, im Layout den Beiträgen zugeordneten und textuell und/oder in Form von Grafiken repräsentierten Metadaten zu den Postings und ihren Produzenten in den Beispielen 14.1–14.6.

Daneben weisen einzelne der in den Abbildungen 14.1–14.6 gezeigten Beispiele auch Merkmale auf, die nicht generell, sondern nur spezifisch für einzelne Kommunikationsumgebungen gelten. Diese Merkmale sollen hier nicht weiter thematisiert werden, da bei der Entwicklung eines Basisschemas zur Repräsentation von Strukturmerkmalen zunächst diejenigen Merkmale im Vordergrund stehen, die übergreifend zu einzelnen Kommunikationsumgebungen eine Rolle spielen. Das schließt nicht aus, dass auch weitere, spezifischere Merkmale in ihm abgebildet werden können oder dass das Schema in einem konkreten Projekt in Hinblick auf die speziellen Erfordernisse der technischen Instanziierung von Interaktion in einzelnen IBK-Umgebungen erweitert werden kann. Der Fokus liegt im Folgenden auf der texttechnologischen Modellierung des Gemeinsamen und seiner Darstellung in Form einer Extension zu den *Guidelines for Text Encoding* der TEI.

4 Ein Basisschema für die Repräsentation internetbasierter Kommunikation in TEI

Ein Basisschema, das für möglichst viele Projekte sinnvoll ist, eine Interoperabilität mit bestehenden Sprachressourcen gewährleistet und flexibel einsetzbar ist, sollte die folgenden Merkmale aufweisen:

1. Es sollte auf etablierte Repräsentationsstandards für Sprachressourcen im Bereich der Korpusliguistik und der Digital Humanities bezogen sein.
2. Es sollte sich für die Bedürfnisse konkreter Korpusprojekte und Forschungsfragen erweitern lassen, ohne die Bedingung (1) zu verletzen.
3. Es sollte so allgemein gehalten sein, dass es als Grundlage für die Annotation in vielen Korpusprojekten nützlich ist.
4. Es sollte so spezifisch sein, dass es grundlegende strukturelle Besonderheiten von IBK-Daten – insbesondere die charakteristischen Unterschiede zur Struktur von redigierten Texten und von Gesprächsverläufen – erfasst.
5. Es sollte in seinen obligatorischen Strukturelementen weitestgehend automatisiert aus gesammelten Rohdaten erzeugt werden können.
6. Es sollte die Erzeugung anonymisierter Sichten auf die Korpusdaten unterstützen.

Anforderung (1) wird im nachfolgend vorgestellten Schema eingelöst durch die Orientierung an den Richtlinien der TEI, die seit 1987 Formate für die Strukturbeschreibung textueller Sprachdaten in den Geisteswissenschaften entwickelt. Die Vorschläge der TEI, die seit 1994 in Form von *Guidelines* vorliegen, können als ein *De-facto*-Standard im Bereich der Geisteswissenschaften gelten, der einer Vielzahl digitaler Forschungsressourcen zugrunde liegt. Der Standard wird von einer breiten Nutzercommunity gepflegt und kontinuierlich weiterentwickelt. Über die Konsistenz der Guidelines wacht ein *Technical Council*. Grundsätzlich kann jeder Interessierte Eingaben für die Weiterentwicklung des Standards machen. Zu Themen, denen die TEI-Community in Hinblick auf die Pflege und Anpassung der Guidelines besondere Relevanz beimisst, sind *Special Interest Groups* (kurz: SIGs) eingerichtet, die sich mit der Nutzung von TEI in einzelnen Anwendungsbereichen oder für bestimmte Textgenres und Kommunikationsbereiche befassen (z. B. *TEI for Linguists*, *Correspondence*, *Manuscripts*, *Scholarly Publishing*).[25] Seit 2013 gibt es eine SIG *Computer-mediated*

25 Die gegenwärtig aktiven SIGs sind auf der Seite http://www.tei-c.org/Activities/SIG/ (letzter Zugriff: 8. 11. 2017) verzeichnet.

Communication (TEI-CMC-SIG), die sich mit der Entwicklung TEI-konformer Formate für Sprachressourcen internetbasierter Kommunikation befasst.

Anforderung (2) wird dadurch gewährleistet, dass TEI die schon erwähnte Möglichkeit der *customization* vorsieht (vgl. Abschnitt 1). Diese erlaubt es, aus dem in den TEI-Richtlinien vorgesehenen Inventar an Modellen (Elementklassen, Elementen, Attributen) Auswahlen zu treffen oder bestimmte Teilinventare zu unterdrücken, neue Elemente und Attribute hinzuzufügen oder bestehende Modelle zu ändern. Solange dabei bestimmte Regeln eingehalten werden, die sicherstellen, dass die vorgenommenen Änderungen sich konsistent in die Architektur des TEI-Frameworks einfügen, stimmt das resultierende Schema grundsätzlich mit den TEI-Richtlinien überein. Für die Erzeugung eigener *customizations* bietet die TEI mit dem Editor *Roma* sogar ein spezialisiertes Tool an.[26]

Da die TEI-Richtlinien bislang keine spezifischen Modelle für die Strukturannotation von IBK-Daten anbieten, arbeitet die TEI-CMC-SIG bei der Entwicklung eines Basisschemas mit der Möglichkeit der *customization*. Bis dato haben Mitglieder der SIG drei Annotationsschemas entwickelt und diese jeweils an Korpora getestet. Der Vorteil der Orientierung an einem Standard und dessen Erweiterung (per *customization*) macht die Schemaentwicklung darüber hinaus ökonomisch: Diejenigen Teile des Schemas, die für die Repräsentation von Korpusdaten benötigt werden, aber keine IBK-spezifischen Merkmale abbilden, können aus dem in TEI schon vorhandenen Standardinventar übernommen werden; Anpassungen und Erweiterungen sind nur da erforderlich, wo der TEI-Standard noch keine geeigneten Lösungen anbietet. Entsprechend stimmen die von der SIG entwickelten Schemas in weiten Teilen mit dem TEI-Standard überein; diejenigen Modelle, die tatsächlich neu eingeführt oder gegenüber dem Standard modifiziert werden mussten, bilden nur einen kleinen Teil.

Die Anforderungen (3), (4) und (5) werden dadurch eingelöst, dass sich das Schema auf die Modellierung der Eckpunkte eines Interaktionsformats beschränkt, das vielen Instanziierungen internetbasierter Kommunikation zugrunde liegt, und dass es sich aus Strukturinformation erzeugen lässt, die in den Korpus-Ausgangsdaten i. a. R. bereits vorhanden ist. Die Anforderung (6) wird eingelöst durch eine Transformation (mindestens) solcher personenbezogener Daten, die sich anhand von Strukturmerkmalen in den Ausgangsdaten identifizieren lassen, in Metadaten.

Die TEI-CMC-SIG hat in einem iterativen Prozess bislang drei Schemas für die Strukturannotation von IBK-Korpora entwickelt. Das erste Schema (Beiß-

26 http://www.tei-c.org/Guidelines/Customization/use_roma.xml (letzter Zugriff: 8. 11. 2017).

wenger et al. 2012; „DeRiK-Schema“[27]) wurde in Zusammenarbeit mit dem Team des DWDS-Projekts an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) als Grundlage für die Repräsentation von IBK-Erweiterungen für Korpora geschriebener Sprache (Beißwenger & Lemnitzer 2013) konzipiert. Das zweite Schema (Chanier et al. 2014; „CoMeRe-Schema“[28]) wurde als Ergebnis verschiedener Diskussionen in der neu gegründeten TEI-CMC-SIG und auf der Grundlage des DeRiK-Schemas von Thierry Chanier und Kollegen für die Repräsentation einer Sammlung von 14 IBK-Korpora zum Französischen entwickelt (CoMeRe-Korpora). Die dritte Schemaversion (CLARIN-D-Schema) entstand im Rahmen des CLARIN-D-Kurationsprojekts *ChatCorpus2CLARIN*, in dem das Dortmunder Chat-Korpus in die Korpusinfrastrukturen am IDS und an der BBAW integriert und in diesem Zusammenhang in TEI remodelliert wurde (Lüngen et al. 2016). Dieses dritte Schema basiert wiederum auf einer Analyse des CoMeRe-Schemas und der mit dessen Anwendung gemachten Annotationserfahrungen und baut das Schema für verschiedene Formen schriftbasierter IBK weiter aus.

Im Folgenden beschreibe ich den letzten Stand des Schemas, das CLARIN-D-Schema.[29] Ich stelle einige wesentliche Merkmale vor, die die in Abschnitt 3 dargestellten Charakteristika schriftbasierter IBK im TEI-Kontext und mit Blick auf die oben formulierten Anforderungen umsetzen. Die bei der Entwicklung des Schemas in das TEI-Rahmenwerk eingebrachten Modifikationen und Erweiterungen sind in einem ODD-Dokument (*One Document Does It All*) beschrieben, das unter derselben Adresse in dem von der TEI vorgegebenen Dokumentationsformat angeboten wird.[30]

In Bezug auf die Repräsentation von IBK-Daten unterscheide ich im Folgenden die folgenden Datentypen und Strukturebenen:[31]

– *Primärdaten:* die eigentlichen Nutzdaten, d. h. diejenigen Sprachdaten und ihre Struktur, an deren Auswertung Nutzer linguistischer Korpora interes-

27 https://wiki.tei-c.org/index.php/SIG:CMC/derikschema (letzter Zugriff: 8. 11. 2017).

28 https://wiki.tei-c.org/index.php?title=SIG:CMC/comereschema (letzter Zugriff: 8. 11. 12017).

29 Das vollständige Schema steht unter https://wiki.tei-c.org/index.php?title=SIG:CMC/clarindschema (letzter Zugriff: 8. 11. 2017) als Schemaspezifikation in der XML-Schemasprache *Relax NG* zur Verfügung und kann in dieser Form für die Korpusannotation eingesetzt werden. An der Entwicklung des Schemas unmittelbar beteiligt waren neben dem Verfasser Axel Herold, Harald Lüngen, Angelika Storrer und Eric Ehrhardt.

30 Zum ODD-Format vgl. Lobin (2010: 110–111).

31 Für die Unterscheidung von Primärdaten, Metadaten und Annotationen greife ich auf eine in der Korpuslinguistik allgemein übliche Differenzierung zurück (vgl. z. B. Storrer 2011: 2018–2219, Perkuhn, Keibel & Kupietz et al. 2012: 45, Lemnitzer & Zinsmeister 2015: 13), die Unterscheidung zweier Strukturebenen in den Primärdaten folgt Beißwenger et al. (2012).

siert sind. In Bezug auf IBK-Daten geht das vorgestellte Schema davon aus, dass es sich bei den Ausgangsdaten, die die Primärdaten für IBK-Korpora bereitstellen, um Dokumente handelt, die eine Abfolge von Postings in einer bestimmten Art der Strukturierung enthalten, die von zwei oder mehreren Autoren produziert wurden (sogenannte *Logfiles* oder Mitschnitte). Strukturen in IBK-Primärdaten lassen sich dabei auf zwei Ebenen beschreiben:

a) auf der Ebene der *Makrostrukturen*: Darunter verstehe ich die spezifische Art der Abfolge und Anordnung von Postings in den Ausgangsdokumenten und im Bildschirmprotokoll der Beteiligten. Makrostrukturen werden charakteristischerweise nicht von einem Autor alleine hergestellt, sondern ergeben sich aus dem Zusammenspiel von Verschickungshandlungen zweier oder mehrerer Autoren *plus* den Aufbereitungs- und Vermittlungsroutinen der zugrunde liegenden Kommunikationstechnologie.
b) auf der Ebene der *Mikrostrukturen*: Darunter verstehe ich die Struktur von Postings und damit all diejenigen Formen der Strukturierung sprachlicher Äußerungen, die der alleinigen Gestaltungshoheit des Autors des Postings unterliegen. Mikrostrukturen lassen sich mit Blick auf die visuelle Gliederung des Posting-Inhalts und die Integration von Medienobjekten beschreiben (Enthält das Posting nur einen oder mehrere Absätze? Enthält es Zwischenüberschriften? Ist mit Listenformatierungen und Einrückungen gearbeitet? Sind Bilder, Audio- oder Videodateien integriert?); Mikrostrukturen können aber auch unter textgrammatischem und syntaktischem Aspekt von Interesse sein (Welche Formen der Verknüpfung von Sätzen nutzt der Autor? Welche syntaktische Struktur weist sein Posting auf?).

- *Metadaten:* Daten, die die Primärdaten beschreiben und die benötigt werden, um die im Korpus dokumentierte Primärdatenstichprobe als solche und auch die darin dokumentierte Form der Sprachverwendung (die im Falle von IBK-Daten interaktional organisiert ist) zu kontextualisieren.
- *Annotationen:* Daten, mit denen Primärdatensegmenten linguistische oder Strukturinformationen zugeordnet werden und die im nachfolgend beschriebenen Basisschema dazu verwendet werden, Strukturinformation auf Makro- und Mikroebene in Form expliziter Beschreibungen zu repräsentieren.

4.1 Metadaten und TEI-Header

Das Schema folgt in seiner Architektur den obligatorischen Strukturvorgaben für TEI-Schemas und interpretiert diese IBK-spezifisch. Standardmäßig umfasst

jedes TEI-Dokument einen *Header*, in dem Metadaten zum Dokument erfasst werden, gefolgt von der Repräsentation der annotierten Daten. Listing 1 zeigt einen Ausschnitt aus der Header-Struktur für ein Dokument aus dem Dortmunder Chat-Korpus. Die Teilstruktur *fileDescription* (Element `fileDesc`) erfasst im *publicationStatement* (Element `publicationStmt`) Angaben zum Anbieter der Ressource (im vorliegenden Fall das IDS und die BBAW) und zu den Lizenzbedingungen, unter denen die Ressource genutzt werden darf (CC BY 4.0) sowie in der *sourceDescription* (Element `sourceDesc`) Angaben zur Herkunft der Daten (in diesem Fall zur Chat-Umgebung, in der das enthaltene Logfile aufgezeichnet wurde, sowie zu Datum und Uhrzeit der Aufzeichnung).[32]

Listing 1: Ausschnitt 1 (gekürzt) aus dem TEI-Header für ein Dokument aus dem Dortmunder Chat-Korpus in CLARIN-D:

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
    <teiHeader>
      <fileDesc>
        <publicationStmt>
          <publisher/>
          <pubPlace/>
          <idno>1204009</idno>
          <distributor>
           CLARIN centre Institut für Deutsche Sprache, Mannheim
           and CLARIN centre Berlin-Brandenburgische Akademie der
           Wissenschaften
           </distributor>
           <date>2016</date>
           <availability>
             <licence target="https://creativecommons.org/ licen
             ses/by/4.0/">CC BY 4.0. Legal restrictions may arise
             from data protection legislation.</licence>
           </availability>
         </publicationStmt>
         <sourceDesc>
          <recordingStmt>
            <recording>
              <equipment>
```

32 Definition und Inhaltsmodell sämtlicher Elemente aus dem TEI-Standard lassen sich in den *Guidelines* nachschlagen (TEI-P5).

```
            <p>plattformName=<name type="OTH">[_CHATPLATFORM
            NAME_]</name></p>
            <p>plattformURL=<ref type="URL">[_WWWURL_]</ref>
            </p>
          </equipment>
          <respStmt>
            <persName xml:id="f1204009.p1">unknown</persName>
            <resp>recording</resp>
          </respStmt>
          <date>2005-01-11</date>
          <time from="22:07:00" to="22:59:00"/>
        </recording>
      </recordingStmt>
...
</fileDesc>
```

Listing 2 zeigt einen weiteren Ausschnitt aus dem TEI-Header. Die Teilstruktur *profileDescription* (Element `profileDesc`) klassifiziert den Dokument-Inhalt mit Referenz auf ein externes Klassifikationsschema. Die Benennung des Elements `textClass` zeigt, dass TEI ursprünglich für die Annotation redigierter, schriftlicher Texte entwickelt wurde, für die die Zuordnung zu einer Textsorte intendiert ist. Im Falle des Chat-Korpus wurde anstelle eines Textklassifikationsschemas eine Klassifikation der enthaltenen Logfiles nach gesellschaftlichen Handlungsbereichen und auf zweiter Klassifikationsebene nach Chat-Plattformen vorgenommen. Unter der im Listing referenzierten URL ist das für das Korpus verwendete Klassifikationsschema zentral hinterlegt. Durch Aufruf der *target*-URL lässt sich entnehmen, dass das im Dokument beschriebene Logfile der Hauptklasse „Professionelle Chats: Chat-Veranstaltungen im Hochschulkontext, Chat-Beratung, Chats im Medienkontext“ sowie innerhalb dieser Klasse der Subklasse „Beratung“ zugeordnet ist und dass das Logfile einem Beratungsangebot mit dem Thema „Beratung durch eBay-Expertin“ entstammt.

Die Teilstruktur *participationDescription* (Element `particDesc`) beschreibt gemäß Definition „the identifiable speakers, voices, or other participants in any kind of text or other persons named or otherwise referred to in a text, edition, or metadata“; für das Chat-Korpus wird das Element für die Hinterlegung von Informationen zu den Chat-Beteiligten adaptiert. Für jeden Beteiligten gibt es in der *listPerson* einen Eintrag mit eindeutiger ID, dem ein Teilnehmername, eine Teilnehmerrolle (hier: *system*, *expert*, *client*) sowie eine Angabe zum (erschlossenen oder vermuteten) biologischen Geschlecht (*sex*) zugeordnet ist. Da das Dortmunder Chat-Korpus für die Integration in CLARIN-D anonymisiert

wurde, sind die ursprünglichen Teilnehmernamen im Listing durch Kategorienlabels ersetzt (zur Anonymisierung vgl. Beißwenger et al. 2017b). Die ID ist wichtig, um im beschriebenen Logfile jedes einzelne Posting per ID-Referenz (vgl. Listing 3, Werte des Attributs *@who*) mit einer eindeutigen Autorenzuordnung versehen zu können. Die Trennung der Angaben zu den Interaktionsbeteiligten vom eigentlichen Logfile und die Referenzierung über IDs hat zwei Vorteile: Zum einen (1) vermeidet sie die redundante Wiederholung immer wieder derselben Beteiligteninformation bei jedem der von diesem Beteiligten produzierten Posting; zum anderen (2) ermöglicht sie bei Korpora, bei denen die Autorennamen nicht anonymisiert wurden, die Erzeugung anonymisierter Sichten, insofern sämtliche Metadaten zu den Beteiligten (hier: Name, Rolle und Geschlecht) einfach vom Logfile abgetrennt werden können; anhand der in den einzelnen Postings angegebenen Beteiligten-IDs ist in diesem Fall eine Unterscheidung der Beteiligten trotzdem noch möglich.

Im Falle des Chat-Logfiles sind die Informationen, die in der *listPerson* für die einzelnen Beteiligten gegeben werden, recht rudimentär. Sie lassen sich aber beliebig ausbauen und erweitern, beispielsweise um Angaben, die aus den Benutzerprofilen der Beteiligten extrahiert wurden (was z. B. im Falle von Online-Foren und Tweets relevant ist), oder um die Inhalte von Benutzersignaturen, die z. B. in Wikipedia-Diskussionen und in Foren den Postings der Beteiligten als automatisch aus Templates generierte Textbausteine beigestellt werden.

Die *timeline* in Listing 2 verzeichnet sämtliche für die Repräsentation des Logfiles relevante Zeitpunktangaben. Diese sind in den Ausgangsdaten den einzelnen Postings typischerweise in Form von sogenannten *Timestamps* (Zeitstempeln) zugeordnet. Die Erfassung als Teil des Headers erlaubt die Repräsentation der zeitlichen Struktur an zentralem Ort und in einem einheitlichen Format, während die Darstellung von Zeitpunktangaben in den *Timestamps* innerhalb desselben Korpus (z. B. in Chat-Logfiles aus verschiedenen Quellen) im Format variieren kann (z. B. „15. Juli 2017, 14:30 Uhr“ vs. „Sonntag, 2:30 p.m.“). Die einzelnen Postings sind den in der *timeline* beschriebenen Zeitpunktangaben (Instanzen des Elements *when*) über ID-Referenzen zugeordnet (vgl. Listing 3, Werte des Attributs *@synch*).

Listing 2: Ausschnitt 2 (gekürzt) aus dem TEI-Header für ein Dokument aus dem Dortmunder Chat-Korpus in CLARIN-D:

```
<profileDesc>
  <textClass>
    <catRef scheme="http://corpora.ids-mannheim.de/
      taxonomies/taxonomy-handlungsbereiche.tei.
      xml#handlungsbereiche"
```

```
        target="http://corpora.ids-mannheim.de/taxonomies/
        taxonomy-handlungsbereiche.tei.xml#h1204000"/>
      </textClass>
      <particDesc>
        <listPerson>
          <person role="system" xml:id="f1204009.A01_System">
            <persName type="nickname">system</persName>
            <sex evidence="estimated">system</sex>
          </person>
          <person role="expert" xml:id="f1204009.A02">
            <persName type="nickname">[_FEMALE-EXPERT-A02_]</
              persName>
            <sex evidence="estimated">female</sex>
          </person>
          <person role="client" xml:id="f1204009.A03">
            <persName type="nickname">[_MALE-CLIENT-A03_]</
              persName>
            <sex evidence="estimated">male</sex>
          </person>
        </listPerson>
      </particDesc>
     <textDesc>
        <interaction>
          <timeline>
            <when absolute="22:07:00" xml:id="f1204009.t001"/>
            <when absolute="22:08:00" xml:id="f1204009.t002"/>
            <when absolute="22:09:00" xml:id="f1204009.t003"/>
            <when absolute="22:10:00" xml:id="f1204009.t004"/>
            <when absolute="22:11:00" xml:id="f1204009.t005"/>
            <when absolute="22:12:00" xml:id="f1204009.t006"/>
            <when absolute="22:13:00" xml:id="f1204009.t007"/>
            <when absolute="22:14:00" xml:id="f1204009.t008"/>
            <when absolute="22:15:00" xml:id="f1204009.t009"/>
            <when absolute="22:16:00" xml:id="f1204009.t010"/>
          …
          </timeline>
        </interaction>
      </textDesc>
    </profileDesc>
</teiHeader>
```

Die Struktur der in den Listings 1 und 2 abgebildeten Ausschnitte aus dem TEI-Header entspricht dem TEI-Standard. Einige Elemente wurden, wie aufgezeigt, allerdings IBK-spezifisch reinterpretiert. Konkrete Änderungen des TEI-Formats gegenüber dem Standard waren auf Ebene der Strukturannotation des eigentlichen Logfiles erforderlich. Hier wurden ganz zentral Lösungen (1) für die Repräsentation von Postings und (2) für die Beschreibung von IBK-Makrostrukturen (Logfiles, Threads) benötigt. Diese Lösungen im CLARIN-D-Schema werden im Folgenden erläutert.

4.2 Annotation von Makro- und Mikrostrukturen

Die TEI-Guidelines umfassen ausgearbeitete Module für die Annotation von Strukturen in redigierten Texten und für die Annotation der Struktur transkribierter Gespräche. In Bezug auf die Frage, wie sich IBK-Strukturen sinnvoll in TEI repräsentieren lassen, ist zunächst zu klären, ob sich diese Strukturen ggf. mit in TEI bereits vorhandenen Modellen sinnvoll darstellen lassen. Drei Modellierungsoptionen aus dem TEI-Standard bieten sich grundsätzlich an und sind zu diskutieren:

- Durch die Brille *redigierter Texte* betrachtet könnten die Primärdaten von IBK-Korpora (die in Dokumenten gespeicherten Logfiles) als *Texte* betrachtet werden, die aus einer Abfolge von Einheiten bestehen, die durch Layoutmerkmale voneinander abgrenzbar sind und von verschiedenen Autoren verfasst wurden. Der Vorteil einer Entscheidung für die Beschreibung von IBK-Strukturen mit den TEI-Modellen für Textstrukturen läge darin, dass für die Beschreibung von IBK-Mikrostrukturen – z. B. die Gliederung komplexer Foren-Postings oder von Einträgen auf Wikipedia-Diskussionsseiten – in TEI ein ausgearbeitetes Inventar an Elementen bereits zur Verfügung stünde. Der Nachteil der damit einhergehenden Beschreibung von IBK-Verläufen als einer *Form von Text* wäre der, dass die Gliederung eines monologischen Textes (in Abschnitte und Absätze, TEI-Elemente `<div>` und `<p>`) typischerweise von *einer* Autor-Instanz konzipiert und verantwortet wird, während es IBK aufgrund ihres interaktionalen Charakters gerade zentral ist, dass die entstehenden Makrostrukturen sich – ähnlich wie in mündlichen Gesprächen – aus dem Zusammenwirken vieler ergeben (unähnlich zu mündlichen Gesprächen spielt, wie oben schon bemerkt, dabei aber auch die Technologie eine nicht unwesentliche Rolle).
- Durch die Brille *mündlicher Gespräche* betrachtet könnten Postings als *utterances* aufgefasst und IBK-Makrostrukturen – analog zu Gesprächs-Transkripten – als Abfolgen von *utterances* (TEI-Element `<u>`) beschrieben werden. Der Vorteil einer solchen Modellierung des Gegenstandes wäre die

Nähe des Modells zur prototypischen Form zwischenmenschlicher Interaktion. Der Nachteil bestünde zum einen darin, dass *utterances* (a) nicht als schriftliche Äußerungen konzipiert sind und (b) dass Postings sich hinsichtlich der für sie charakteristischen, konsekutiven Abfolge von Verbalisierung, Übermittlung und adressatenseitiger Verarbeitung (vgl. Abschnitt 3) fundamental anders verhalten als interaktionale mündliche Äußerungen. Das Konzept der *utterance* müsste so umfassend reinterpretiert werden, dass es für eine distinktive Erfassung von Turns in mündlicher Interaktion nicht mehr brauchbar wäre. Ebenso wie die Möglichkeit der Beschreibung von Postings als Textgliederungen (Textabschnitte oder Absätze) bedeutete eine Beschreibung von Postings als *utterances* eine unzulässige Verschleierung zentraler Charakteristika der Interaktionskonstitution in IBK.

– Durch die Brille des TEI-Modells für *performance texts*, das für die Strukturannotation von Dramentexten konzipiert ist, könnten Postings als *speeches* (TEI-Element <sp>) aufgefasst werden: <sp> beschreibt „an individual speech in a performance text, or a passage presented as such in a prose or verse text". Tatsächlich ist das Element <sp> das einzige TEI-Element, das für die Repräsentation interaktionaler Äußerungen vorgesehen ist, die primär (und nicht wie *utterances* in transkribierten Gesprächen erst durch sekundäre Verschriftung) medial schriftlich realisiert sind. Auch dieses Element scheidet aber aus naheliegenden Gründen für die Repräsentation von IBK-Strukturen aus: *speeches* beschreiben keine authentischen Sprachäußerungen und die durch *speeches* konstituierten Makrostrukturen sind – unter Produktionsaspekt monologisch – von einer einzigen Autor-Instanz konzipiert.

Im vorgestellten Schema wie auch schon in den beiden Vorgängerschemas haben wir uns daher dafür entschieden, für die Annotation von Postings per *customization* ein eigenes TEI-Element <post> einzuführen und dieses mit einem Inhaltsmodell auszustatten, das sowohl die Gemeinsamkeiten, als auch die charakteristische Differenz einerseits zu Gliederungseinheiten in monologischen Texten – also zu Einheiten, die im TEI-Universum üblicherweise als oder <p> repräsentiert werden – als auch andererseits zu Sprecherbeiträgen in mündlichen Interaktionen – in TEI mit <u> dargestellt – zur Geltung bringt. Eventuelle Gemeinsamkeiten oder Unterschiede zu *speeches* werden nicht weiter verfolgt.

Das Element <post> und die durch Abfolgen von <post>-Instanzen konstituierten IBK-Makrostrukturen weisen die folgenden Merkmale auf:

– <post> ist konzipiert als ein Element vom Typ *model.divPart*. Eine Anforderung bei der *customization* besteht darin, dass neu hinzugefügte, im

Standard nicht enthaltene Elemente in das Elementklassensystem des TEI-Rahmenwerks eingeordnet werden müssen. Damit wird beschrieben, wie sich ein neu hinzugefügtes Element zu bereits vorhandenen Elementen und ihren Inhaltsmodellen verhält, welche Kind-Elemente und Attribute es von vorhandenen Modellen erbt. Durch die Konzeption von <post> als *model.divPart* wird ausgesagt, dass Postings in formaler Hinsicht Ähnlichkeiten mit Einheiten der Textgliederung aufweisen: In einem gespeicherten Logfile sind sie typischerweise als Gliederungseinheiten zu erkennen; anhand von Layouteigenschaften lassen sich Logfiles damit weitgehend automatisiert in Einheiten des Typs <post> zergliedern (vgl. die o. a. Anforderung (5)).

- Zugleich wird über das Attribut *@who* die Möglichkeit geschaffen, jedem Posting einen individuellen Autor zuzuordnen. Damit wird dargestellt, dass die IBK-Makrostruktur, die sich als Abfolge von Postings darstellt, nicht das Ergebnis einer von einer einzelnen Autor-Instanz verantworteten, monologischen Strukturbildung ist; stattdessen wird sie als interaktionale Struktur konzipiert. Dieses Merkmal teilen IBK-Makrostrukturen, vermittelt über die Konzeption des Elements <post>, mit mündlichen Gesprächen.
- Die Makrostruktur als solche wird durch das Element aus dem TEI-Standard dargestellt. kann Textgliederungen unterschiedlichster Art beschreiben. In den TEI-CMC-Schemas wird reinterpretiert als eine Gliederungseinheit, die Sequenzen aus zwei oder mehr Instanzen des Elements <post> bündelt. Elemente des Typs lassen sich über ein Attribut *@type* subklassifizieren, für das als Wertbelegungen *logfile* und *thread* empfohlen werden, um verschiedene Typen von IBK-Makrostrukturen zu unterscheiden.
- Das Element <post> lässt sich über eine Reihe von neu eingeführten oder für die Repräsentation von IBK adaptierten Attributen subklassifizieren (u. a.):
  a) Mit *@type* werden verschiedene Typen von Postings unterschieden: „Standard"-Postings in direkter Rede sowie Postings vom Typ „event", mit denen – insbesondere in Chats – eine Aussage aus „Regisseurssicht" über reale oder spielerisch vollzogene Aktivitäten eines Interaktionsbeteiligten formuliert wird („FrankieABC holt sich mal nen Kaffee", „FrankieABC betritt den Raum").
  b) *@replyTo* ermöglicht (optional) eine weitergehende Beschreibung der sequenziellen Vernetzung des Postings mit anderen Postings durch Verweis auf die entsprechenden Posting-IDs. Die Realisierung von *ReplyTo*-Beziehungen in der Annotation ist sowohl automatisiert als auch manuell denkbar: automatisiert durch automatische Auswertung

von Adressierungselementen oder Zitationen in der Mikrostruktur der Postings (vorausgesetzt, die entsprechenden Elemente sind auf der Ebene der Mikrostruktur ihrerseits annotiert); manuell als Ergebnis einer qualitativen Analyse von Postings auf dem Hintergrund ihres sequenziellen Kontexts.

c) *@auto* gibt an, ob ein Posting automatisch generiert wurde. Typische Fälle sind sogenannte „Systemmeldungen" in Chats, mit denen Resultate von Nutzereingaben, die nicht unmittelbar als Postings intendiert sind, gemeldet werden (z. B. „Friede23 betritt den Raum"). Für quantitative Korpusuntersuchungen ist es wichtig, solche Postings bei der Korpusanalyse ausfiltern zu können, da ihre Inhalte nicht auf Verbalisierungen von menschlichen Interaktionsbeteiligten zurückgehen, sondern aus von den Programmierern im System vordefinierten Templates erzeugt wurden. Das muss nicht nur auf Posting-Ebene der Fall sein; auch auf der Mikroebene von Postings können Textbausteine dieser Art vorkommen, beispielsweise in Einleitungen zu Zitatblöcken in Online-Forenbeiträgen, die über die automatische Zitierfunktion erzeugt wurden.

Listing 3 zeigt einen Ausschnitt aus einem Korpusdokument mit drei vollständig annotierten `<post>`-Instanzen, die in einem `<div>`-Element vom Typ *logfile* gebündelt sind. Die Werte zu den Attributen *@who* und *@synch* referenzieren die IDs von Einträgen in der *listPerson* und in der *timeline* aus dem TEI-Header (vgl. Listing 2 und zugehörige Erläuterungen). Auf der Mikroebene der drei abgebildeten Postings ist zudem das Ergebnis einer automatischen *Part-of-speech*-Annotation und einer Lemmatisierung, die mit den in Horbach et al. (2014) beschriebenen Sprachverarbeitungswerkzeugen erzeugt wurden, in Form einer Inline-Annotation in die TEI-Repräsentation integriert. Dazu wurde das Element `<w>` (*word*) aus dem TEI-Standard als Element für die Beschreibung morphosyntaktischer Informationen adaptiert. Die *Part-of-speech*-Informationen folgen dem für IBK-Korpora erweiterten STTS-Tagset nach Beißwenger et al. (2015).

Listing 3: Ausschnitt aus dem TEI-Body für ein Dokument aus dem Dortmunder Chat-Korpus in CLARIN-D:

```
<div type="logfile">
    <post auto="false" rend="color:black" synch="#f1204009.t040"
    type="standard" who="#f1204009.A02" xml:id="f1204009.m187">
        <time> 22:46 </time>
```

```
    <anchor type="sentence_start"/>
    <w lemma="ich" type="PPER" xml:id="f1204009.m187.
        t1">Ich</w>
    <w lemma="wünschen" type="VVFIN" xml:id="f1204009.m187.
        t2">wünsche</w>
    <w lemma="Sie|sie" type="PPER" xml:id="f1204009.m187.
        t3">Ihnen</w>
    <w lemma="alle" type="PIAT" xml:id="f1204009.m187.
        t4">alles</w>
    <w lemma="Gute" type="NN" xml:id="f1204009.m187.
        t5">Gute</w>
    <w lemma="für" type="APPR" xml:id="f1204009.m187.
        t6">für</w>
    <w lemma="ihr" type="PPOSAT" xml:id="f1204009.m187.
        t7">Ihre</w>
    <w lemma="Verhandlung" type="NN" xml:id="f1204009.m187.
        t8">Verhandlungen</w>
    <w lemma="mit" type="APPR" xml:id="f1204009.m187.
        t9">mit</w>
    <w lemma="die" type="ART" xml:id="f1204009.m187.
        t10">dem</w>
    <w lemma="DPD" type="NN" xml:id="f1204009.m187.
        t11">DPD</w>
</post>
<post auto="false" rend="color:black" synch="#f1204009.t040"
type="standard" who="#f1204009.A02" xml:id="f1204009.m188">
    <time> 22:46 </time>
    <anchor type="sentence_start"/>
    <w lemma="und" type="KON" xml:id="f1204009.m188.
        t1">und</w>
    <w lemma="wünschen" type="VVFIN" xml:id="f1204009.m188.
        t2">wünsche</w>
    <w lemma="Sie|sie" type="PPER" xml:id="f1204009.m188.
        t3">Ihnen</w>
    <w lemma="alle" type="PIAT" xml:id="f1204009.m188.
        t4">alles</w>
    <w lemma="Gute" type="NN" xml:id="f1204009.m188.
        t5">Gute</w>
    <w lemma="für" type="APPR" xml:id="f1204009.m188.
        t6">für</w>
```

```
        <w lemma="die" type="ART" xml:id="f1204009.m188.
            t7">den</w>
        <w lemma="zukünftig" type="ADJA" xml:id="f1204009.m188.
            t8">zukünftigen</w>
        <w type="NN" xml:id="f1204009.m188.
            t9">Online-Handel</w>
        <w lemma="!" type="$." xml:id="f1204009.m188.
            t10">!</w>
    </post>
    <post auto="false" rend="color:black" synch="#f1204009.t040"
    type="standard" who="#f1204009.A03" xml:id="f1204009.m189">
        <time> 22:46 </time>
        <anchor type="sentence_start"/>
        <w lemma="ja" type="PTKANT" xml:id="f1204009.m189.
            t1">Ja</w>
        <w lemma="," type="$," xml:id="f1204009.m189.
            t2">,</w>
        <w lemma="die" type="PDS" xml:id="f1204009.m189.
            t3">das</w>
        <w lemma="können" type="VMFIN" xml:id="f1204009.m189.
            t4">kann</w>
        <w lemma="ich" type="PPER" xml:id="f1204009.m189.
            t5">ich</w>
        <w lemma="gebrauchen" type="VVINF" xml:id="f1204009.
            m189.t6">gebrauchen</w>
        <w type="EMOASC" xml:id="f1204009.m189.t7">:-(</w>
    </post>
…
</div>
```

Im Projekt *ChatCorpus2CLARIN* wurde das vorgestellte Schema für die Remodellierung des kompletten Dortmunder Chat-Korpus (1 Million Tokens) in TEI angewendet (Lüngen et al. 2016). Neben Chat-Daten flossen in die Entwicklung des Schemas Analysen und Annotationsexperimente zu einem Datenset mit Stichproben weiterer IBK-Formen ein (Wikipedia-Diskussionsseiten, WhatsApp-Interaktionen, Newskommunikation, Tweets). Die Vorgängerversion von Beißwenger et al. (2012) wurde von Margaretha & Lüngen (2014) für die Repräsentation der Posting-Struktur im Wikipedia-Diskussionsseiten-Korpus in DeReKo adaptiert. Die Vorgängerversion von Chanier et al. (2014) lag der Strukturannotation von vierzehn französischen IBK-Korpora zu neun unter-

schiedlichen IBK-Formen zugrunde (u. a. SMS, Wikipedia-Diskussionen, Tweets, Weblogs, E-Mails, Foren, Chats). Die aktuelle Schemaversion wird gegenwärtig an der Universität Gießen für die Strukturannotation eines Scienceblog-Korpus eingesetzt (Grumt Suárez, Karlova-Bourbonus & Lobin 2016). In den genannten Projekten hat sich das Basisschema als praktikabel erwiesen. Im Rahmen von *ChatCorpus2CLARIN* hat die Verwendung des Schemas eine Integration der Ressource in vorhandene Korpussammlungen zur geschriebenen deutschen Sprache ermöglicht, die bereits in Standard-TEI repräsentiert sind (DeReKo, DWDS).

5 Fazit und Ausblick

Die Sprachressourcen-Infrastruktur der Zukunft wird IBK-Korpora umfassen, die
- eine breite Zahl von IBK-Genres und Sprachen abdecken,
- frei für Forschung und Lehre zur Verfügung stehen, getreu dem Motto der europäischen Sprachressourcen-Infrastruktur-Initiative CLARIN „Sprachressourcen für alle“ (*Common Language Resources*),
- für linguistische Analysezwecke aufbereitet, d. h. um linguistische Strukturannotationen angereichert und durch Metadaten erschlossen sind,
- in Übereinstimmung mit Standards im Bereich der Digital Humanities repräsentiert sind,
- interoperabel sowohl untereinander als auch mit Korpora anderen Typs (Textkorpora, Korpora gesprochener Sprache) sind und sich daher mit denselben Korpusrecherchewerkzeugen vergleichend mit anderen Korpora auswerten lassen.

Diese Vision ist nicht unrealistisch: In einer wachsenden Zahl von Projekten zu unterschiedlichen Sprachen entstehen derzeit Korpora zur internetbasierten Kommunikation und werden Lösungen für die mit diesem Korpustyp verbundenen Desiderate entwickelt. Standardisierungs- und Infrastrukturinitiativen im Bereich der Digital Humanities haben den Bedarf an Forschungsressourcen zur internetbasierten Kommunikation erkannt und unterstützen die Entwicklung von Lösungen, um sie in die Ressourcenlandschaft einzugliedern und mit existierenden Ressourcen zu vernetzen.

Ein wichtiger Baustein beim Aufbau interoperabler IBK-Korpora ist ein Standard für die Repräsentation von IBK-Daten. Die TEI bietet ein flexibles Rahmenwerk, um einen solchen Standard zu entwickeln. Die von der TEI-CMC-SIG vorgelegten Entwürfe stehen als vollständige TEI-Schemas zur Verfügung, wurden begleitend zu ihrer Entwicklung in verschiedenen Korpusprojekten er-

probt und können von jedermann für die Annotation eigener Korpora genutzt werden. Nach wie vor handelt es sich bei diesen Schemas um *customizations*: Sie sind mit dem TEI-Standard vollumfänglich kompatibel, enthalten aber dennoch einzelne Modelle, die selbst nicht Teil des Standards sind. Ein wichtiger nächster Schritt in der Arbeit der CMC-SIG, in der Akteure von Korpusprojekten zu unterschiedlichen Sprachen mitwirken, besteht daher darin, aus den bislang vorgelegten *customizations* und den mit ihrer Anwendung gemachten Erfahrungen Eingaben für den TEI-Standardisierungsprozess zu formulieren.

Ein Standard für die Repräsentation von IBK muss sich notwendigerweise auf die Bereitstellung von Modellen beschränken, die basale Interaktionsformate darstellbar machen, die in vielen Kommunikationsumgebungen im Internet eine Rolle spielen und die sich als relativ stabil gegenüber dem beständigen technologischen Wandel erweisen. Das in diesem Beitrag beschriebene Posting-basierte Interaktionsformat, das zahlreichen Formen internetbasierter Kommunikation zugrundeliegt, ist ein Beispiel für ein solches basales Format. In einem weiteren Schritt wurde gezeigt, wie sich dieses Format mit dem aktuellen Schemaentwurf der TEI-CMC-SIG in einer XML-Repräsentation darstellen lässt, die mit dem TEI-Standard kompatibel ist. Eine künftige Aufgabe der CMC-SIG wird es sein, dieses Basisformat, das vor allem auf schriftbasierte Formen internetbasierter Kommunikation fokussiert, um Modelle für die Repräsentation von Daten aus multimodalen IBK-Umgebungen zu erweitern. Erste Ansätze dafür, wenn auch im vorliegenden Beitrag nicht näher behandelt, sind im CoMeRe-Schema (Chanier et al. 2014) angelegt und wurden im CLARIN-D-Schema fortgeschrieben. Ein drittes wichtiges Desiderat stellt die Anpassung von Schemata für die Erfassung von Metadaten für den Bereich der IBK dar: Gerade weil sich Kommunikationstechnologien und -umgebungen beständig wandeln, müssen Beschreibungen zur Struktur und zum Funktionsumfang von Kommunikationsumgebungen, aus denen Daten erhoben worden, in einer Weise erfasst und repräsentiert werden, dass Nutzer der entsprechenden Korpora in 10 oder 20 Jahren, möglicherweise auch schon in 3 Jahren, noch rekonstruieren können, welche technologischen Rahmenbedingungen die sprachliche Gestaltung von Interaktion, so wie es sich in den auch Daten zeigt, beeinflusst haben.

Ein Standard für die Repräsentation und Strukturannotation stellt dabei letzten Endes nur *einen* Baustein für die Verbesserung der Ressourcenlage in Bezug auf die Domäne der internetbasierten Kommunikation dar. Um die als Vision skizzierte IBK-Korpuslandschaft der Zukunft zu realisieren, müssen Innovationen bei der sprachtechnologischen Verarbeitung von IBK-Daten und bei der Anpassung von Werkzeugen für die Korpusrecherche und -analyse, *best practices* für die Adressierung juristischer und forschungsethischer Fragen bei der Datenerhebung und -dokumentation sowie Lösungen für die Anonymisierung der Korpusdaten als weitere Bausteine hinzutreten.

Die Korpora internetbasierter Kommunikation von morgen bilden einen (möglicherweise bedeutsamen) Teil der kulturellen Überlieferung des Alltagssprachgebrauchs von heute für die Sprachhistoriker von übermorgen.[33] Die daraus resultierenden Anforderungen an die Bereitstellung und Repräsentation von Korpora ist damit nicht nur gegenwartsbezogen, sprachen- und domänenübergreifend, sondern auch diachron bezogen eine wichtige Aufgabe bei der Arbeit an der Sprachressourceninfrastruktur der Zukunft.

Literatur

Auer, Peter (2000): On line-Syntax – oder: was es bedeuten könnte, die Zeitlichkeit der mündlichen Sprache ernst zu nehmen. In *Sprache und Literatur* 85, 43–56.

Barbaresi, Adrien (2016): Efficient construction of metadata-enhanced web corpora. In *Proceedings of the 10th Web as Corpus Workshop, Association for Computational Linguistics*, 7–16. https://hal.archives-ouvertes.fr/hal-01371704v2/document (letzter Zugriff: 8. 11. 2017).

Barbaresi, Adrien & Kay-Michael Würzner (2014): For a fistful of blogs: Discovery and comparative benchmarking of republishable German content. In *Proceedings of NLP4CMC workshop (KONVENS 2014)*, 2–10. Hildesheim University Press. https://www.dwds.de/static/publications/pdf/Barbaresi-Wuerzner_Fistful-of-blogs_2014.pdf (letzter Zugriff: 8. 11. 2017).

Beißwenger, Michael (2003): Sprachhandlungskoordination im Chat. *Zeitschrift für germanistische Linguistik* 31 (2), 198–231.

Beißwenger, Michael (2007): *Sprachhandlungskoordination in der Chat-Kommunikation*. Berlin, New York: de Gruyter (Reihe Linguistik – Impulse & Tendenzen 26).

Beißwenger, Michael (2010): Chattern unter die Finger geschaut: Formulieren und Revidieren bei der schriftlichen Verbalisierung in synchroner internetbasierter Kommunikation. In Vilmos Ágel & Mathilde Hennig (Hrsg.), *Nähe und Distanz im Kontext variationslinguistischer Forschung*, 247–294. Berlin, New York: de Gruyter.

Beißwenger, Michael (2013): Das Dortmunder Chat-Korpus. In *Zeitschrift für germanistische Linguistik* 41 (1), 161–164.

Beißwenger, Michael (2016): Praktiken in der internetbasierten Kommunikation. In Arnulf Deppermann, Helmuth Feilke & Angelika Linke (Hrsg.), *Sprachliche und kommunikative Praktiken*. Jahrbuch 2015 des Instituts für Deutsche Sprache, 279–310. Berlin/New York: de Gruyter.

Beißwenger, Michael (Hrsg.) (2017): *Empirische Erforschung internetbasierter Kommunikation*. Berlin, New York: de Gruyter (Empirische Linguistik/Empirical Linguistics 9).

33 Dieser Gedanke geht auf einen Diskussionsbeitrag von Erhard Hinrichs im Rahmen eines deutsch-französischen Kolloquiums zu Standards für IBK-Korpora an der Universität Duisburg-Essen zurück (Gedächtniszitat). Ich danke Erhard Hinrichs für diese weitsichtige Anregung.

Beißwenger, Michael & Angelika Storrer (2008): Corpora of computer-mediated communication. In Anke Lüdeling & Merja Kytö (Hrsg.), *Corpus Linguistics HSK* 29 (1), 292–309. Berlin: de Gruyter.

Beißwenger, Michael & Angelika Storrer (2012): Interaktionsorientiertes Schreiben und interaktive Lesespiele in der Chat-Kommunikation. *Zeitschrift für Literaturwissenschaft und Linguistik* 168, 92–124.

Beißwenger, Michael & Lothar Lemnitzer (2013): Aufbau eines Referenzkorpus zur deutschsprachigen internetbasierten Kommunikation als Zusatzkomponente für die Korpora im Projekt „Digitales Wörterbuch der deutschen Sprache" (DWDS). *Journal for Language Technology and Computational Linguistics* 28 (2), 1–22.

Beißwenger, Michael, Nelleke Oostdijk, Angelika Storrer & Henk van den Heuvel (2014): Building and Annotating Corpora of Computer-Mediated Communication: Issues and Challenges at the Interface of Corpus and Computational Linguistics. Journal of Language Technology and Computational Linguistics 2. http://jlcl.org (letzter Zugriff: 8. 11. 2017).

Beißwenger, Michael, Thomas Bartz, Angelika Storrer & Swantje Westpfahl (2015): Tagset und Richtlinie für das Part-of-Speech-Tagging von Sprachdaten aus Genres internetbasierter Kommunikation. Guideline-Dokument aus dem Projekt „GSCL Shared Task: Automatic Linguistic Annotation of Computer-Mediated Communication/Social Media" (EmpiriST 2015). http://ttps//sites.google.com/site/empirist2015/home/annotation-guidelines (letzter Zugriff: 8. 11. 2017).

Beißwenger, Michael, Maria Ermakova, Alexander Geyken, Lothar Lemnitzer & Angelika Storrer (2012): A TEI Schema for the Representation of Computer-mediated Communication. *Journal* of *the Text Encoding Initiative* 3. http://jtei.revues.org/476 (doi:10.4000/jtei.476) (letzter Zugriff: 8. 11. 2017).

Beißwenger, Michael, Thierry Chanier, Tomaž Erjavec, Darja Fišer, Axel Herold, Nikola Lubešic, Harald Lüngen, Céline Poudat, Egon Stemle, Angelika Storrer & Ciara Wigham (2017a): Closing a Gap in the Language Resources Landscape: Groundwork and Best Practices from Projects on Computer-mediated Communication in four European Countries. *Selected Papers from the CLARIN Annual Conference 2016*, October 26–28 2016, France, Aix-en-Provence. Linköping: Linköping University Electronic Press (Linköping University Electronic Conference Proceedings), 1–18.

Beißwenger, Michael, Harald Lüngen, Jan Schallaböck, John H. Weitzmann, Axel Herold, Pawel Kamocki, Angelika Storrer & Julia Wildgans (2017b): Rechtliche Bedingungen für die Bereitstellung eines Chat-Korpus in CLARIN-D: Ergebnisse eines Rechtsgutachtens. In Michael Beißwenger (Hrsg.), *Empirische Erforschung internetbasierter Kommunikation*. Berlin, New York: de Gruyter (Empirische Linguistik/Empirical Linguistics 9), 7–46.

Bolander, Brook & Miriam A. Locher (2014): Doing Sociolinguistic Research on Computer-Mediated Data: A Review of Four Methodological Issues. *Discourse, Context & Media* (3), 14–26.

Brinker, Klaus, Hermann Cölfen & Steffen Pappert (2014): *Linguistische Textanalyse: eine Einführung in Grundbegriffe und Methoden*. 8., neu bearb. und erw. Aufl. Berlin: Erich Schmidt.

Chanier, Thierry, Celine Poudat, Benoit Sagot, Georges Antoniadis, Ciara Wigham, Linda Hriba, Julien Longhi & Djamé Seddah (2014): The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. Journal of language Technology and

Computational Linguistics 29 (2), 1–30. http://www.jlcl.org/2014_Heft2/1Chanier-et-al.pdf (letzter Zugriff: 8. 11. 2017).

Cherny, Lynn (1999): *Conversation and Community. Chat in a Virtual World*. Stanford: University of Chicago Press.

Chiari, Isabella & Alessio Canzonetti (2014): Le forme della comunicazione mediata dal computer: generi, tipi e standard di annotazione. In Enrico Garavelli & Elina Suomela-Härmä (Hrsg.), *Dal manoscritto al web: canali e modalità di transmissione dell'italiano. Tecniche, materiali e usi nella storia della lingua*. Atti del XII Convegno della Società Internazionale di Linguistica e Filologia Italiana (SILFI), 18–19 June 2012, Helsinki, 595–606. Florenz: Franco Cesati Editore.

CLARIN-D AP 5 (2012): *CLARIN-D User Guide*. Version: 1.0.1. https://www.clarin-d.de/en/help/user-handbook (letzter Zugriff: 8. 11. 2017).

Deppermann, Arnulf, Helmuth Feilke & Angelika Linke (Hrsg.) (2016): *Sprachliche und kommunikative Praktiken*. Jahrbuch 2015 des Instituts für Deutsche Sprache. Berlin/New York: de Gruyter.

DiDi (2015): Beschreibung der Anonymisierung im DiDi-Korpus. http://www.eurac.edu/en/research/autonomies/commul/Documents/DiDi/DiDi_anonymisation_DE.pdf (letzter Zugriff: 8. 11. 2017).

Dürscheid, Christa (2005): Medien, Kommunikationsformen, kommunikative Gattungen. *Linguistik online* 22 (1). http://www.linguistik-online.de/22_05/duerscheid.pdf (letzter Zugriff: 8. 11. 2017).

Dürscheid, Christa & Elisabeth Stark (2011): sms4science: An international corpus-based texting project and the specific challenges for multilingual Switzerland. In Crispin Thurlow & Kristine Mroczek (Hrsg.), *Digital Discourse. Language in the New Media*, 299–320. Oxford: Oxford University Press.

Ehlich, Konrad (1984): Zum Textbegriff. In Annely Rothkegel & Barbara Sandig (Hrsg.), *Text – Textsorten – Semantik*, 531–550. Hamburg: Buske.

Fišer, Darja, Tomaž Erjavec & Nikola Ljubešić (2016): JANES v0.4: Korpus slovenskih spletnih uporabniških vsebin. *Slovenščina* 2.0 4(2), 67–99. http://dx.doi.org/10.4312/slo2.0.2016.2.67-99 (letzter Zugriff: 8. 11. 2017).

Fišer, Darja & Michael Beißwenger (Hrsg.) (2017): *Investigating Computer-Mediated Communication: Corpus-Based Approaches to Language in the Digital World*. Ljubljana: Ljubljana University Press (Translation Studies and Applied Linguistics). Open Access: https://knjigarna.ff.uni-lj.si/en/izdelek/1766/investigating-computer-mediated-communication/

Fišer, Darja, Tomaž Erjavec & Nikola Ljubešić (2017): The compilation, processing and analysis of the Janes corpus of Slovene user-generated content: In Ciara R Wigham & Gudrun Ledegen (Hrsg.), *Corpus de Communication Médiée par les Réseaux. Construction, structuration, analyse*, 125–138. Paris: L'Harmattan (Humanités numériques).

Forsyth, Eric N. & Craig H. Martell (2007): Lexical and Discourse Analysis of Online Chat Dialog. In *Proceedings of the First IEEE International Conference on Semantic Computing (ICSC 2007)*, USA, Irvine, 19–26, https://catalog.ldc.upenn.edu/docs/LDC2010T05/lex_anal_online_chat.pdf (letzter Zugriff: 8. 11. 2017).

Frey, Jennifer-Carmen, Egon W. Stemle & Aivars Glaznieks (2014): Collecting Language Data of Non-Public Social Media Profiles. In Gertrud Faaß & Josef Ruppenfofer (Hrsg.), *Workshop Proceedings of the 12th Edition of the KONVENS Conferenc*e, 11–15. Hildesheim: Universitätsverlag Hildesheim.

Frey, Jennifer-Carmen, Aivars Glaznieks & Egon W. Stemle (2016): The DiDi Corpus of South Tyrolean CMC Data: A Multilingual Corpus of Facebook Texts. In *Proceedings of the Third Italian Conference on Computational Linguistics* (CLiC-it 2016), 5–6 December 2016, Italy, Napoli, http://ceur-ws.org/Vol-1749/paper27.pdf (letzter Zugriff: 8. 11. 2017).

Garcia, Angela Cora & Jennifer Baker Jacobs (1998): The Interactional Organization of Computer Mediated Communication in the College Classroom. *Qualitative Sociology* 21 (3), 299–317.

Garcia, Angela Cora & Jennifer Baker Jacobs (1999): The Eyes of the Beholder: Understanding the Turn-Taking System in Qua-si-Synchronous Computer-Mediated Communication. *Research on Language and Social Interaction* 32 (4), 337–367.

Geyken, Alexander, Adrien Barbaresi, Jörg Didakowski, Bryan Jurish, Frank Wiegand & Lothar Lemnitzer (2017): Die Korpusplattform des „Digitalen Wörterbuchs der deutschen Sprache“ (DWDS). *Zeitschrift für germanistische Linguistik* 45 (2), 327–344.

Giesbrecht, Eugenie & Stefan Evert (2009): Is Part-of-Speech Tagging a Solved Task? An Evaluation of POS Taggers for the German Web as Corpus. In *Proceedings of the 5th Web as Corpus Workshop (WAC5)*. San Sebastian, Spain. http://www.stefan-evert.de/PUB/GiesbrechtEvert2009_Tagging.pdf (letzter Zugriff: 8. 11. 2017).

Grumt Suárez, Holger, Natali Karlova-Bourbonus & Henning Lobin (2016): Compilation and Annotation of the Discourse-structured Blog Corpus for German. *In Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanitie*s (cmc-corpora2016), University of Ljubljana. http://nl.ijs.si/janes/wp-content/uploads/2016/09/CMC-2016_Grumt_et_al_Compilation-and-Annotation.pdf (letzter Zugriff: 8. 11. 2017).

Herring, Susan C. (1996): *Computer-Mediated Communication. Linguistic, Social and Cross-Cultural Perspectives*. Amsterdam, Philadelphia: John Benjamins Publishing Company (Pragmatics & Beyond New Series 39).

Herring, Susan C. (1999): Interactional Coherence in CMC. *Journal of Computer-Mediated Communication* 4 (4), doi:10.1111/j.1083–6101.1999.tb00106.x (letzter Zugriff: 8. 11. 2017).

Horbach, Andrea, Diana Steffen, Stefan Thater & Manfred Pinkal (2014): Improving the Performance of Standard Part-of-Speech Taggers for Computer-Mediated Communication. In *Proceedings of KONVENS 2014*, 171–177. Hildesheim: Universitätsverlag Hildesheim.

Horsmann, Tobias & Torsten Zesch (2015): Effectiveness of Domain Adaptation Approaches for Social Media PoS Tagging. In *Proceeding of the Second Italian Conference on Computational Linguistics*, 166–170. Trento: Accademia University Press.

Imo, Wolfgang (2015a): Vom ikonischen über einen indexikalischen zu einem symbolischen Ausdruck? Eine konstruktionsgrammatische Analyse des Emoticons :-) In Kerstin Fischer, Anatol Stefanowitsch, Alexander Lasch & Jörg Bücker (Hrsg.), *Konstruktionsgrammatik 5. Konstruktionen im Spannungsfeld von sequenziellen Mustern, kommunikativen Gattungen und Textsorten*, 133–162. Tübingen: Stauffenburg-Verlag.

Imo, Wolfgang (2015b): Vom Happen zum Häppchen … Die Präferenz für inkrementelle Äußerungsproduktion in internetbasierten Messengerdiensten. *Networx* 69, http://www.mediensprache.net/de/networx/networx-69.aspx (letzter Zugriff: 8. 11. 2017).

iRights.Law Rechtsanwälte (2016): *Rechtsgutachten zur Integration mehrerer Text-Korpora in die CLARIN-D-Infrastrukturen*. (Manuskript, 46 Seiten).

JIM-Studie (2016): Jugend, Information, (Multi-)Media. Basisuntersuchung zum Medienumgang 12–19-Jähriger. Hrsg. v. Medienpädagogischen Forschungsverbund Südwest. http://www.mpfs.de/de/studien/jim-studien/2016/

Jucker, Andreas H. & Christa Dürscheid (2012). The Linguistics of Keyboard-to-screen Communication. A New Terminological Framework. *Linguistik Online* 56, 39–64.

König, Katharina (2015): Dialogkonstitution und Sequenzmuster in der SMS- und WhatsApp-Kommunikation. *Travaux neuchâtelois de linguistique* 63, 87–107.

Lemnitzer, Lothar & Heike Zinsmeister (2015): *Korpuslinguistik. Eine Einführung*. 3., überarb. u. rew. Aufl. Tübingen: Narr (Narr Studienbücher).

Lindemann, Katrin, Emanuel Ruoss & Caroline Weinzinger (2014): Dialogizität und sequenzielle Verdichtung in der Forenkommunikation: Editieren als kommunikatives Verfahren. *Zeitschrift für Germanistische Linguistik* 42 (2), 223–252.

Ljubešić, Nikola, Darja Fišer, Tomaž Erjavec, Jaka Čibej, Dafne Marko, Senja Pollak & Iza Škrjanec (2015): Predicting the level of text standardness in user-generated content. In *Proceedings of the 10th International Conference on Recent Advances in Natural Language Processing*, Bulgaria, Hissar, 371–378.

Lobin, Henning (2010): *Computerlinguistik und Texttechnologie*. München: Fink.

Lüngen, Harald & C. M. Sperberg-McQueen (2012): A TEI P5 Document Grammar for the IDS Text Model. *Journal of the Text Encoding Initiative (jTEI)* 3, http://jtei.revues.org/508 (letzter Zugriff: 8. 11. 2017).

Lüngen, Harald, Michael Beißwenger, Axel Herold & Angelika Storrer (2016): Integrating corpora of computer-mediated communication in CLARIN-D: Results from the curation project ChatCorpus2CLARIN. In Stefanie Dipper, Friedrich Neubarth & Heike Zinsmeister (Hrsg.), *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, 1561–64.

Lüngen, Harald (2017): DeReKo – Das Deutsche Referenzkorpus. Schriftkorpora der deutschen Gegenwartssprache am Institut für Deutsche Sprache in Mannheim. *Zeitschrift für germanistische Linguistik* 45 (1), 161–170.

Margaretha, Eliza & Harald Lüngen (2014): Building Linguistic Corpora from Wikipedia Articles and Discussions. *Journal of language Technology and Computational Linguistics* 29 (2), 59–82. http://www.jlcl.org/2014_Heft2/3MargarethaLuengen.pdf (letzter Zugriff: 8. 11. 2017).

Markman, Kris (2006): *Computer-Mediated Conversation: The Organization of Talk in Chat-Based Virtual Team Meetings*. Dissertation. University Texas at Austin.

Marx, Konstanze (2015): „kümmert euch doch um euren Dreck" – Verteidigungsstrategien im Cybermobbing dargestellt an einem Beispiel der Plattform Isharegossip.com. In U. Tuomarla et al. (Hrsg.), *Misskommunikation und Gewalt. Mémoires de la Société Néophilologique de Helsinki*, 125–138. Vantaa: Hansaprint Oy.

Murray, Denise E. (1989): When the medium determines turns: turn-taking in computer conversation. In Hywel Coleman (Hrsg.), *Working with Language. A Multidisciplinary consideration of Language Use in Work Contexts*, 319–337. Berlin, New York: de Gruyter Mouton.

Oostdijk, Nelleke, Martin Reynaert, Véronique Hoste & Ineke Schuurman (2013): The Construction of a 500 Million Word Reference Corpus of Contemporary Written Dutch. In Peter Spyns & Jan Odijk (Hrsg.), *Essential Speech and Language Technology for Dutch: Results by the STEVIN-programme*, 219–247. Berlin: Springer Verlag.

Perkuhn, Reiner, Holger Keibel & Marc Kupietz (2012): *Korpuslinguistik*. München: Fink.

Schönfeldt, Juliane & Andrea Golato (2003): Repair in Chats: A Conversation Analytic Approach. *Research on Language and Social Interaction* 36 (3), 241–284.

Schröck, Jasmin & Harald Lüngen (2015): Building and Annotating a Corpus of German-Language Newsgroups. In *Proceedings of the 2nd Workshop on Natural Language*

Processing for Computer-Mediated Communication/Social Media (NLP4CMC2015). Germany, Essen, 17–22. https://sites.google.com/site/nlp4 cmc2015/program (letzter Zugriff: 8. 11. 2017).

Selting, Margaret & Elizabeth Couper-Kuhlen (2000): Argumente für die Entwicklung einer interaktionalen Linguistik. *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion* 1, 76–95. http://www.gespraechsforschung-ozs.de (letzter Zugriff: 8. 11. 2017).

Severinson Eklundh, Kerstin (2010): To Quote or Not to Quote: Setting the Context for Computer-Mediated Dialogues. *Language@Internet* 7. http://www.languageatinternet.org/articles/2010/2665 (letzter Zugriff: 8. 11. 2017).

Stertkamp, Wolf (2016): *Spiel, Satz, Sieg: Sprache und Kommunikation in Online-Computerspielen. Eine qualitative Analyse multimodaler Kommunikation in Massively Multiplayer Online Role-Playing Games am Beispiel von Word of Warcraft*. Dissertation. Justus-Liebig-Universität Gießen.

Storrer, Angelika (2001): Sprachliche Besonderheiten getippter Gespräche: Sprecherwechsel und sprachliches Zeigen in der Chat-Kommunikation. In Michael Beißwenger (Hrsg.), *Chat-Kommunikation. Sprache, Interaktion, Sozialität und Identität in synchroner computervermittelter Kommunikation. Perspektiven auf ein interdisziplinäres Forschungsfeld*, 3–24. Stuttgart: ibidem.

Storrer, Angelika (2011): Korpusgestützte Sprachanalyse in Lexikographie und Phraseologie. In Karlfried Knapp (Hrsg.), *Angewandte Linguistik*. 3., vollst. überarb. und erw. Aufl., 216–239. Tübingen: Francke.

Storrer, Angelika (2014): Sprachverfall durch internetbasierte Kommunikation? Linguistische Erklärungsansätze – empirische Befunde. In Albrecht Plewina & Andreas Witt (Hrsg.), *Sprachverfall? Dynamik – Wandel – Variation. Jahrbuch des Instituts für Deutsche Sprache 2013*, 171–196. Berlin, Boston: de Gruyter.

Storrer, Angelika (2018): Interaktionsorientiertes Schreiben im Internet. In: Deppermann, Arnulf (Hg.): Sprache im kommunikativen, interaktiven und kulturellen Kontext, 219–244. Berlin, New York: de Gruyter.

TEI Consortium (2007): *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. http://www.tei-c.org/Guidelines/P5/ (letzter Zugriff: 8. 11. 2017).

Verheijen, Lieke & Wessel Stoop (2016): Collecting Facebook Posts and WhatsApp Chats. In *Proceedings. Text, Speech, and Dialogue: 19th International Conference*, Czech Republic, Brno, September 12–16, 249–58. Cham: Springer International Publishing.

[WAC-X/EmpiriST 2016] *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*. Stroudsburg: Association for Computational Linguistics (ACL Anthology W16–26). http://aclweb.org/anthology/W/W16/W16-26.pdf (letzter Zugriff: 8. 11. 2017).

Wigham, Ciara R. & Gudrun Ledegen (Hrsg.) (2017): *Corpus de Communication Médiée par les Réseaux*. Paris: L'Harmattan.